THE VALIDITY OF REPUTATION SURVEY BASED ACADEMIC RANKING SYSTEMS FOR U.S. DOCTORAL PROGRAMS

A THESIS
SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE
MASTER OF SCIENCE

By

MATTHEW T. BRENNEMAN

WITH DR. REBECCA PIERCE AS ADVISOR

BALL STATE UNIVERSITY

MUNCIE, INDIANA

DECEMBER 2016

December 2016

ACKNOWLEDGMENTS

Any research project requires a lot of work and is rarely done alone in a vacuum. I feel very fortunate to have been in a department where I was given the freedom and latitude to pursue my own intellectual path but always had a "safety net" when I needed help. First, I want to sincerely thank my thesis advisor, Professor Rebecca Pierce for her support and help throughout my entire graduate career at Ball State University. Professor Pierce not only helped inspire and guide my thesis project, but she played a crucial role in helping me gain a deep appreciation for statistics and was the main reason I changed career paths from research to statistics education. I am profoundly grateful for all of the time, advice, and assistance she so selflessly gave me. I also want to thank both of my other committee members, Professor Munni Begum and Professor Rahmatullah Imon for all of their help and advice. I learned a great deal from Professor Begum about both statistics and teaching. Despite her full schedule, Professor Begum always found time to "talk shop" with me, and I greatly benefitted from her good research sense and enjoyed her enthusiasm and deep appreciation for intellectual discovery. Professor Imon was also very kind and generous with his time, and I always appreciated his considered suggestions and ideas.

I also wish to thank the men and women of the staff in the Mathematical Sciences department who were always there to help me. Thank you to Susan Bourne and Carol Deiwert for all of their administrative assistance (and making sure I was paid), and to Mr. Joel Bozell for his copier expertise (and great sense of humor). I wish especially to thank Professor John Lorch for all his support, understanding, and help in my graduate teaching assistant duties. Finally, I want to thank the Department of Mathematical Sciences for their financial support without which this thesis would never have been written.

Table of Contents

List of Tables

Section 1

List of Figures

SECTION 1 INTRODUCTION

1.1 An Overview and Critique of Academic Ranking Systems (ARS)

The first known ARS was published by the American psychologist James McKeen Cattell in 1910 (Cattell, 1910). Cattell had a keen interest in what he called "merit ranking". Using a list he compiled of 5,500 eminent American scientists, Cattell ranked psychology departments based on the number of faculty at each institution that were on his list. Donald Hughes, a Miami University of Ohio chemistry professor, improved on this idea by using surveys based on peer reputation to rank graduate programs (Hughes, 1925). Even though a few other academic ranking systems for schools and graduate programs were produced between 1936 and 1970, they did not enjoy a wide circle of popularity with the general public. Students instead tended to use publications that did not rank schools, but rather gave a profile and information about individual programs (such as the Peterson Guide series) and colleges (like Barron's Guide to Colleges). Rankings of academic programs and institutions of higher learning did not become popular until The U.S. News & World Report (USNWR) published their first ranking of U.S. colleges and universities in 1983. The USNWR ranking system became popular with the public because it came at a time when both college enrollment and educational costs were increasingly rapidly. It filled the need of many students to have a data based method for making decisions about schools that was simple and easy to understand. Since the inception of the USNWR ratings, many new systems (or league tables, as they are called in Europe) have been developed by a number of private, media based, and educational organizations all over the world. As the diversity of student needs and educational programs has increased, the number of ARS has likewise proliferated to the point where there are

well over 100 such systems ("College and University Rankings", 2016). Most of the academic ranking systems and related research on them pertain to the ranking of academic institutions (globally and regionally) and professional programs (such as law, medicine, engineering, and business) with relatively lesser attention given to graduate programs.

It is interesting that despite the large number of academic ranking systems produced and all the studies that have been performed on them, no methodology for testing their validity has been proposed. From the viewpoint of predictive modeling, this state of affairs is due to foundational problems in the definition of the response variable (i.e. "academic quality"), the choice of covariates, and a test of a given ARS's validity (both internally and externally). One way to study model validity is to classify the ARS as to how they estimate academic quality. In this sense, we can divide ARS into two classes: those that measure academic quality using only reputational surveys ("reputation survey based ARS") and those that define academic quality as a weighted average of different measures of an academic system's performance ("weighted ARS").

Weighted ARS make up the majority of ARS. All of the major recognized methods for globally ranking academic institutions, for example, use this approach. It is useful to express the mathematical formula for the ranking of a given academic system in the following general form:

$$r \overset{defn}{=} \sum_{i=1}^{n} \alpha_i y_i \qquad (1)$$

$$y_i \overset{defn}{=} \sum_{j=1}^{m_i} \beta_{ij} x_{ij} \qquad (2)$$

where r is the response variable called a "rank score" that measures a system's "academic quality". As equation (1) indicates, the rank score is a weighted linear combination of numerical measures for the system's performance in the different areas (where y_i is the i^{th} area of performance). There are a large number of potential areas of an academic system's performance that could be used,

(such as teaching, research, various student outcomes, and diversity), but no rational basis for which areas are used in any given ARS. Most weighted ARS use those metrics that reflect the qualities they believe constitute a quality education. An ARS for undergraduate institutions therefore might use areas such as teaching, cost, and student outcomes, while an ARS for a graduate program might heavily weight research. The measures of academic performance in each area, as equation (2) shows, are themselves a weighted linear combination of quantities that measure that quality (x_{ij} being the j^{th} metric used for y_i). There are generally many possible metrics that can be used to measure the quality of a particular area, which again, generally reflects the purpose of those that create the ARS. For example, "quality of teaching" might be measured by class size for a system that ranks U.S. undergraduate institutions, while the number of Nobel Laureates might be used for a system that ranks the top universities globally.

The three most common ways to compute the response variable, r, for a weighted ARS are to use a weighted average of the standardized y_i values, a weighted average of the percentiles for the y_i (where the weights are positive and sum to unity) or to use an "input-output" ranking system. An input-output ranking system is restricted to a specific collection of academic systems where the "input" (generally capital expenditures of some sort) and "output" (some measure of academic productivity, such as number of publications) for each school is computed. The total output and input for all the schools is computed which is used to compute and rank the difference in the percentage of the total output and percentage of the total input for each school (Kivinen and Hedman, 2008).

However irrespective of how the ranking is computed, the fundamental flaw in all weighted ARS is that the response variable, r, cannot be measured. A consequence of this fact is there is no rational basis for the process of model selection, making the choice of which areas should be

considered in estimating academic quality, (i.e. the y_i), and their respective weights, (i.e. the α_i), largely subjective. Added to this is the fact that the choice of metrics used to measure the performance in each area (i.e. the x_{ij}) and their weights (i.e. the β_{ij}) are also subjective. Table 1 illustrates how the subjective nature of this method leads to a large number of different global weighted ARS that purport to measure the same thing.

Table 1: Measures of Academic Performance Used By the Major Global Weighted ARS

ARS (reference)	Areas of Academic Performance (y_i)	Metrics to Estimate Performance in Academic Area (x_{ij})
U.S. News and World Report ("How U.S. News Calculated", 2016)	Research quality	Peer reputation surveys
	Publications	Number of publications, citations, and books
	Global Connections	International collaborations
	Production of doctorates	Number of PhDs awarded and number of PhDs per academic staff member
Academic Ranking of World Universities ("Academic Ranking of World Universities", 2016)	Quality of education	Number of alumni that are Nobel Laureates and Fields Medalists
	Faculty quality	Number of faculty that are Nobel Laureates, Fields medalists or highly cited
	Publications	Number of publications that are highly cited or in journals with high impact factors
Webometrics ("Ranking web of universities", 2016)	Web presence	Number of online publications, visitors, and web pages
Center for World University Rankings ("Center for World University Rankings", 2016)	Educational Quality	Per capita awards, prizes, and medals won by alumni
	Alumni employment	Per capita alumni holding CEO positions at top companies
	Publications	Number of publications, citations, and impact factors
	Faculty quality	Number of awards, medals, and prizes won and patents filed
CWTS Leiden Ranking ("CWTS Leiden Rankings: Indicators", 2016)	Publications	Number of publications, citation, impact factors, collaborations, and global collaborations

A second serious consequence of the fact the response variable cannot be measured is there is no way to assess if a model is valid and the scope of its validity. This is an issue because given the large diversity of academic institutions, it is unreasonable to expect the same weightings to be

appropriate for all schools or programs (i.e. the areas and weights used for a small liberal arts teaching college will not be the same as those for a large research university).

Putting the subjective nature of the weighting schemes aside, another weakness of weighted ARS is that they are susceptible to fraud and manipulation. The transparency of weighted ARS, initially supported by academic institutions, was eventually used by schools to fabricate, manipulate, and manufacture data to improve their ranking. In 1995, a Wall Street Journal article (Stecklow, 1995) reported that many schools had manipulated SAT scores to improve their rankings. For example, some schools dropped a certain percentage of the SAT scores (either the lowest scoring students or those in certain groups, such as remedial or international students) to increase their average SAT score and therefore improve their rankings. Worse yet were schools that reported data which was accurate but had been manufactured solely for the sake of the rankings. In 2008, the New York Times reported that Baylor University had offered students financial incentives to retake their SAT exams (Rimer, 2008). Other means by which schools have "gamed the system" include soliciting applications of students they don't intend to accept to decrease the percentage of applicants accepted (Tierney, 2013), jury rigging class sizes so the percentage of classes with less than 20 students increases, and increasing graduation rates via grade inflation and watered down curriculum (Lederman, 2009).

Overall, weighted academic ranking systems have little credibility among researchers and administrators. Alan J. Stone, president of Alma College, described the arbitrary selection and weighting of the metrics used in weighted ARS as being 'so subjective, it is ridiculous.' (Webster and Mare, 2005). Although the overwhelming majority of school administrators share his opinion that the rankings are not useful, the pressure to participate is so strong that to date only an estimated 5% of the U.S. colleges and universities have opted out (Diver, 2005).

The other method used in academic ranking systems is reputation surveys. These surveys are given to school administrators and experienced faculty who are asked to rank the academic quality of similar institutions or programs. Survey methods are perhaps the most commonly used ranking methodology of graduate programs. Although reputational surveys do not involve the subjective choice of sub measures and weights, the question of whether they are valid measures of academic quality is still an issue. Initially, the results for reputational surveys were generally accepted because they are based on the well-accepted principle of academic peer review. Upon closer scrutiny, however, serious questions regarding validity and bias have surfaced. First, experience has shown that expert opinion is not perfect. Systems based solely on expert opinion, like the Delphi Method, often have built into them cycles so experts may reassess their opinion after hearing those of other experts. This concern certainly has merit in the case of reputation survey based ARS. While years ago, there were not that many doctoral programs and the academic community in a field was often small, the number of doctoral programs has increased into the hundreds for most disciplines. A study that evaluated how knowledgeable respondents in an academic peer survey study of 158 U.S liberal arts schools found that 84% of those surveyed were unfamiliar with many of their peer institutions (Liu and Cheng, 2005). Second, surveys of almost any type generally contain some sources of bias. The same study by Liu and Cheng showed one fourth of these respondents who were not familiar with a peer academic system just guessed. Although one might think such guessing would in a sense average out, there is evidence to suggest that in the absence of an informed opinion on an academic system, respondents will tend to be biased by the institution's reputation or the presence of a star researcher. Findings from studies involving reputational surveys of graduate programs suggest that this so called "halo effect" may exist and serve to bias rankings in favor of well-known established schools. A study of perhaps

the best-regarded graduate ARS, conducted by the NRC, shows that roughly 85% of the variance in the assessed scholarly quality of graduate program faculty is accounted for by institution size and factors related to undergraduate admissions selectivity (Grunig, 1997). This finding is supported by the fact that there tends to be large positive correlations among the USNWR graduate program rankings and their undergraduate rankings (Austin, 1985). Other types of bias, not as well studied, that could also potentially affect reputational survey results are negative bias due to administrators at rival schools and the positive bias of those who graduated from the school they are rating (Lawrence and Green, 1980).

The arbitrary nature of weighted ARS and questions about the validity of reputational survey based ARS has led many to conclude that ARS are not meaningful. Moreover, there is fear that ARS will influence the direction of academics by forcing institutions to care more about looking good on paper than offering a quality education. Many argue that since intellectual discovery plays a major role in the development of major advances, this loss of focus on the part of universities will adversely affect our society and make academia irrelevant to those outside of it (Shapinker, 2008). As a consequence, many groups have called for a moratorium on such ranking systems in order to analyze them and develop a method that better assesses scholarly contribution (Adler and Harzing, 2009).

1.2 Overview of Graduate Programs Academic Ranking Systems (GPARS)

Graduate programs were one of the first academic systems to be ranked in 1925 by Donald Hughes using reputational surveys (Hughes, 1925). Most of the major graduate program ARS created after Hughes were also based on reputational surveys (Cartter, 1966; Keniston, 1959; National Science Foundation, 1969). Later, methodologies for GPARS were proposed based on a single measure of academic production (Clark, 1957) and weighted ARS (Clark, 1974).

In contrast to general academic institutions, for which many ARS have been created, only four ranking systems for graduate programs are known. The methodologies used in the four recognized GPARS are shown in Table 2.

Since the ranking systems for graduate programs and institutions use the same general methodology, they are subject to the same general methodological problems and criticisms. However, unlike institutional rankings, that are largely based on weighted ARS, the two main GPARS, the NRC and USNWR, are based on reputational surveys. There is however a more vocal constituency that is strongly opposed to reputational surveys and more in favor of using measures of productivity for GPARS (Davis and Diamond, 1997).

Although the NRC study is not without its problems and critics, it is perhaps the most comprehensive, balanced, and well-conducted GPARS. It is based on a census of graduate programs taken every 5 years by the NRC, and strives to provide a more balanced view of graduate rankings. The NRC compiles a great many statistics about each school, which they advise should accompany their ratings (as opposed to just a list of ranked schools). Unlike the USNWR, the NRC makes its data available to the public, and prides itself on transparency and careful data collection. Another significant aspect of the NRC GPARS is their responsiveness to feedback in trying to improve their results.

Table 2: Comparison of Main GPARS	
GPARS	**Ranking methodology**
Graduate Programs.com	Survey of students who are asked to rank programs in the following 15 areas: • Academic Competitiveness • Affordability & Campus Safety • Career Support • Educational Quality • Faculty Accessibility & Support • Use of technology • Social Life • Student Diversity • Surrounding Areas • Transportation • Quality of Network • Financial Aid • Grad Program Value • Workload
PhD.org	Combination of peer surveys in 20 different areas. Computes scores in 4 areas and allows users to adjust their weightings • Overall program quality (based on NRC survey of 20 factors) • Research productivity ("composite measure of research productivity, based on publications per faculty member, citations per publication, percent of faculty holding grants, and awards per faculty member") • Student outcome (composite score reflecting time to degree, employment after graduation, graduation rates, and % students with full financial support, and whether school collects employment data on graduates) • Percentage of faculty that are tenured
NRC	Based on a census of graduate programs taken every 5 years Does not give a ranking but a range of rankings in 5 areas • S-Rank: Faculty are asked to give weights to 20 characteristics and then these weights were used to compute range of rankings • R-Rank: Faculty were asked to rank PhD programs near them and the regression was used to determine which factors went into this decision. Using this model, ranges for program rankings was determined. • Research: based on data such as faculty publications, citations, grants, and awards • Student based: students' completion rates, financial aid, and other criteria. • Diversity: gender balance, ethnic diversity, and proportion of international students.
USNWR	Sends surveys to faculty in program being rated to schools that have graduated at least 5 PhDs in that field in the last 5 years. Respondents are asked to rate all schools on list on a 0-5 scale. Average scores are computed and the schools ranked accordingly

Perhaps the most important aspect of the NRC GPARS, however, is that it does propose a rational basis for the construction of weighted ARS. The NRC addresses both the problem of

model selection and issues of the scope for a model's validity by asking experts to take two surveys. One survey is called "the S-rank survey" and the other is called "an R rank survey". The S-rank survey asks each respondent to choose (from a fixed list of 20 items) which factors they feel are most important. The R-rank survey is a reputational rating of peer programs. Taking the rankings (from the R-rank) as the response variable and the factors from the S-survey as the potential covariates, a regression analysis is performed to determine the weights on each factor. Moreover, because programs are grouped based on what factors are chosen on the S-survey, they take into account differences in mission and purpose graduate programs may have. Another unique feature of the NRC rankings is they provide a 90% confidence interval for their rankings through resampling methods. This is an important point because when a large number of schools are ranked (especially on a fixed scale), the difference in ranking scores of two neighboring schools may give the appearance that the ARS is incredibly precise. When in fact, it is well known from a number of studies that outside of the top ranked schools, the ranking of schools near the bottom is not robust (Webster, et al, 1991). However, at the very heart of the NRC GPARS lies the untested assumption that peer reputation rankings provide a valid measure of a graduate program's quality. As discussed in the next section, this consideration is not hypothetical and it is one of the major reasons some researchers have concluded reputational surveys should be eliminated from GPARS completely.

1.3 The Problem of Testing the Validity of Reputation Survey Based GPARS

ARS in general, pose an interesting problem from a statistical point of view, because they raise the question as to the limitations of statistical inference. In the absence of a testable and valid model, many have taken the default position that educational quality is something that "counts but cannot be counted". This is expressed in the words of Dr. Colin Diver, past President of Reed College, who said:

"*Higher education is not a mass-produced commodity but an artisan-produced, interactive, and individually tailored service of remarkable complexity. Trying to rank institutions of higher education is a little like trying to rank religions or philosophies. The entire enterprise is flawed, not only in detail but also in conception*" (Diver, 2005).

However, how valid is this claim that academic quality is something that cannot be quantified and therefore ranked? A cursory glance at most ratings seem to indicate a portion of the results seem reasonable and comport with many of our expectations: the well-known prestigious schools, for example, tend to be at the top of the list. This gives credence to the notion many academic rankings do seem to be measuring something meaningful (i.e. they have some validity as measures of educational quality). However, the fact that significant variations exist between any two ARS shows the overall validity of ARS is a major problem, which must be resolved. Without some way of determining if an ARS is actually measuring the academic quality of an academic system, little confidence can be placed in any ARS.

By elimination, the only methodology that is a potential candidate for being statistically valid is reputation survey based GPARS. In principle, it is reasonable reputation surveys would be a good candidate for a valid ARS since it is rooted in the peer-review principle that scientific, scholarly, and artistic quality is best assessed by recognized experts in the field. This principle has

been found to be very robust and has enjoyed wide respect among not only academics, but also government, business, and foundation officials as the most appropriate method for awarding appointments, promotions, tenure, research grants, contracts, and prizes. Rather than simply accepting the validity of reputational surveys by default, it is argued peer review actually addresses President Diver's argument that academic systems cannot be ranked because "Higher education is not a mass-produced commodity but an artisan-produced, interactive, and individually tailored service of remarkable complexity". Although this statement may appeal personally to academicians, who perhaps view academics as something sacred, on a rational basis Diver's argument does not hold water. His argument is deconstructed in three steps to show why peer review (at least in principle) is well suited for ranking academic systems. The first step is to realize if Diver's argument is to be taken at face value, then we have to conclude Reed College itself has no means by which it can judge how successful it is in meeting its own goals. If the educational outcomes Reed College has set for itself cannot be quantified, then the only way the president can judge how successful the college is in meeting those goals is through the use of anecdotal evidence and his or her own limited experiences. Not only does it not make sense the head of any major organization does not use statistics to evaluate its performance, but Diver himself later in the article specifically points to several metrics which Reed feels are important (such as number of students going on to earn PhDs, the number of honor theses earned, etc.). Although Diver opines that many factors put into the system (such as the hours a professor puts into helping a student on an honors thesis) are difficult to quantify, the outcomes (which are the entities are being ranked) surely are not. The second point is to note the only way Diver's position makes sense logically is if we are to believe Reed College is so unique it cannot be compared to other institutions. While Reed College does occupy a special niche as a small, rigorous, liberal arts college, there are many

institutions that fit this criterion (such as Oberlin, Grinnell, Carleton, etc.). Therefore, we conclude there are institutions with which Reed College can be compared, which segues into the last and final observation. The same criteria by which Reed judges the quality of its programs will generally be the same criteria by which those institutions in its class will also judge the quality of their own institutions. This observation makes clear that peer review by administrators and faculty at comparable institutions, can in principle, be valid measures of academic quality. In short, this argument gives a rational basis for the belief peer review provides a holistic way to judge and rank academic systems quantitatively.

This thesis looks specifically at one of the major arguments against the validity of reputational surveys: the halo effect. This problem will be considered specifically in the context of the 2013 USNWR rankings of U.S. statistics graduate programs. Before directly addressing this problem, it is first important to consider other sources of bias and show they will not significantly affect this study. First, personal bias is a factor which must be considered, since the fact some of these institutions are competitors could potentially affect the survey results. In this study, however, it will be assumed that personal biases will not be a major concern. Competition notwithstanding, the experts are acting in a professional capacity and if they harbored such a significant amount of personal bias, then we would have to conclude that other peer review systems (such as those for research publications) were also dysfunctional. Collective opinion also generally tends to reduce the effect of biased responses and the fact that each respondent is allowed to rate their own institution will act to counter balance negative ratings. Second, respondents who are not sufficiently knowledgeable to rate other peer academic systems must also be considered. The basis for this concern was a previous study that showed a significant proportion of respondents were ignorant of other academic systems and some of these individuals will make up a rating (Liu and

Chen, 2005). Since the survey designed by USNWR however gives respondents the option of not ranking a program if they do not feel qualified to (with "NR", for no response), this effect may not have a significant effect on the rankings considered in this study. The one major remaining problem therefore is the influence the "halo effect" may have on the reputational graduate program ratings.

Although many studies have suggested the presence of a "halo effect" by showing significant correlations between program specific and non-program specific measures of academic quality, a major criticism of this approach is such correlations may be perfectly natural since high quality graduate programs tend to reside in high quality institutions. One premise for a statistical test is if reputational surveys are valid then it would follow program specific metrics of academic quality should better fit reputational surveys than non-program specific metrics. Using this notion, a statistical hypothesis test is built by partitioning the metrics upon which the reputational surveys are thought to depend into two classes: program specific (p) and non-program specific (~p). The null hypothesis would be the non-program specific metrics account for as much variance or more than the program specific metrics, which could be expressed in terms of a formal hypothesis test on the adjusted R^2 value as:

$$H_0 : \left(R^2_{adj}\right)_p \geq \left(R^2_{adj}\right)_{\sim p}$$
$$H_1 : \left(R^2_{adj}\right)_p < \left(R^2_{adj}\right)_{\sim p}$$

SECTION 2 METHODOLOGY

2.1 Overview of Purpose, Context, and Process of Data Collection

This thesis focuses on the application of a statistical hypothesis test for the validity of the reputational survey based ranking of U.S. statistics doctoral programs performed by USNWR. The study presented here is preliminary and therefore incomplete: only a number of the possible variables that could be considered were actually analyzed (owing mostly to time limitations). The basic idea underlying the test is to collect two different sets of data on each institution. One set of data involves metrics that reflect the general quality of the program's home institution (which we shall call "non-program specific metrics") while the other set of data contains metrics that directly reflect the quality of the statistics doctoral program (which we will call "program specific metrics"). The USNWR rank score for each institution is regressed separately against each set of metrics, program specific and non-program specific. Validity is judged from the results of a hypothesis test on the adjusted R^2 value for the two different sets of metrics.

Which metrics are used is largely a matter of choice, but previous studies give some indications as to which are good candidates for modeling. Broadly speaking, the main non-program specific metrics past studies indicate are significantly correlated with graduate program rankings are standardized test scores, institutional control (private or public), admissions selectivity, and endowment. Program specific metrics considered significant with respect to graduate program rankings are publication volume, number of faculty, number of doctoral degrees awarded, and research funding. The specific metrics used, the sources, and details regarding their context are

presented in Table 3. Non-program and program specific metrics are shown in red or blue colored type, respectivley.

Table 3: Specifics Regarding Data Collected on Performance Metrics Studied

Metric	Source (reference)	Context
Admissions factors	IPEDS ("IPEDS Data Center",2016)	Data on a number of factors was collected such as admissions policy, secondary school GPA, secondary school rank, secondary school record, completion of college prep program, and recommendations
Institutional Control	IPEDS	Binary data (Public or Private)
Endowment	IPEDS and Wikipedia site for each school (along with the websites for some schools)	Initial endowment values were obtained from IPEDS. Due to the financial structure of some institutions, the numbers do not accurately represent the endowment for schools considered. Therefore, these values were crossed checked on Wikipedia, and when the values from the two schools differed by more than $10 million dollars, the websites of individual schools were researched.
Percent Accepted	IPEDS	Percent of Fall 2014 applicants who were accepted
Percent Admissions Yield	IPEDS	Percentage of applicants accepted full time for Fall 2014 that accepted.
MATH SAT Scores	IPEDS	First and third quartile scores
Critical Reading SAT Scores	IPEDS	First and third quartile scores
Web Ranking	Webometrics ("Ranking web of universities", 2016)	Obtained the world rank of each institution and their relative rank
Number of Publications	Web of Science	Number of publications in statistics for each school (see section 3 for further details).
Number of PhDs Awarded	NSF-WebCASPAR ("WebCASPAR", 2016)	Searched NSF database using the CIP codes for biostatistics (26.1102) and general statistics (27.501) from 2008 to 2012.
Number of Faculty	Website of each program listed	From website for each program, counted the number of potential research faculty (by excluding adjuncts, instructors, professor emeriti, and teaching faculty)
Faculty to PhD awarded ratio	From two data sources given above	Divided number of PhDs awarded from 2008 to 2012 by number of potential research faculty

The data for the response variable, the rank score for each statistics graduate program, was obtained from the USNWR website for the 2013 statistics graduate program rankings ("Best Statistics Programs", 2016) along with some additional material and information regarding the methodology used in computing the scores from Ms. Angela Pitts, an analyst at USNWR.

USNWR ranks non-professional graduate programs, such as statistics, periodically (roughly every four years). The American Statistical Association (ASA) at the request of USNWR provided the population of schools initially considered for ranking. The ASA compiled this list based on the instructions from USNWR to include any school that had produced 5 PhDs in statistics over the 4-year period from 2009 to 2012. The original list of schools sent by the ASA to USNWR that satisfy this criterion is given in Appendix 1[1]. There are 87 schools on this list. A cursory glance reveals that some schools are listed twice (such as UC Berkeley, Stanford University, University of Pennsylvania and the University of Washington) because some universities have two statistics related departments that are being ranked (such as a statistics and a biostatistics department). From this list, it was found there were 43 statistics programs, 31 biostatistics/epidemiology programs, and 3 business statistics/decision science/OR programs. This information was crosschecked by going to the website for each program. One of the programs (the statistics program at Case Western Reserve University) had been temporarily suspended, and was therefore taken off the list.

The survey methodology outline by USNWR ("Methodology: Best Science Schools Rankings", 2016) states that two surveys were sent to each school on the list: one to the head of the department and a second to a senior faculty member (generally, the graduate program advisor). An identical copy of the survey was sent to every respondent. A copy of this survey with the following instructions is shown in Appendix 2[2]:

[1] Information provided by Jeffrey Myers, Public Relations coordinator ASA, Personal Communication
[2] Provided by Ms. Angela Pitts, analyst for USNWR

Please rate the academic quality of the doctoral statistics program at each school with which you are familiar. Consider all factors that bear on or give evidence of the excellence of the school's doctoral statistics program, for example, curriculum, record of scholarship, quality of faculty and graduates.

The ranking scale used in the survey is also included as shown below in Figure 1.

Figure 1: Rating Scale Used in USNWR Survey

Out of the 174 surveys distributed to all of the schools on the ASA list, a response rate of 39% was obtained [3]. To insure the number of responses for each school was sufficiently large to estimate the rankings well, USNWR required each school to have at least 10 responses that were not "NR". USNWR felt too few schools from the 2012 survey satisfied this criterion, thus, they combined the rankings of the statistics programs they had obtained in 2008 with those they had obtained in 2012[4]. The ranking for each school was computed according to the formula given in equation 4:

$$Rank\,Score = \frac{Sum\ of\ all\ scores\ from\ 2008 + Sum\ of\ all\ scores\ from\ 2012}{Total\ number\ of\ respondents\ that\ ranked\ school\ from\ 2008\ and\ 2012} \quad (4).$$

[3] Personal communication, Ms. Angela Pitts, analyst at USNWR
[4] Personal communication, Ms. Angela Pitts, analyst at USNWR

In order for a program to appear on their published list, a program had to have a rank score of 2.0 or higher. US News and World Report would not provide the raw total scores, nor would they provide any data on those schools which were listed but not ranked (because their score was below 2). Their policy is to provide this information upon request of a school administrator. The final published list of programs with their rank score is given in Appendix 3.

2.2 General Methodology for Data Preparation and Analysis

The rank scores for the schools in the 2013 published USNWR of statistics graduate programs were copied from the USNWR website. There were originally 79 programs ranked. Not all of these 79 programs were used in our analysis. Seven of these programs were removed from the list because they were biostatistics departments that resided primarily in medical schools. There is evidence (Zhang and Chen, 2015) that faculty ratings in these environments are largely dependent on NIH and NSF funding. These two variables are not included in this study. In addition, some of these seven programs were housed in medical schools (such as Medical College of Wisconsin, Medical University of South Carolina, and University of Texas Health Science Center) and were not directly affiliated with an academic institution having an undergraduate program, and hence could not be used in this analysis. For these two reasons, these seven programs were dropped from the list. Two other programs, the NYU Stern School of Business and Kansas State University, were also removed because they did not supply adequate data to National Center for Educational Statistics on important variables used in this study (such as SAT scores). Finally, the statistics program at Case Western Reserve University was dropped because its statistics program was in the process of merging with the mathematics department and the statistics graduate program was temporarily suspended.

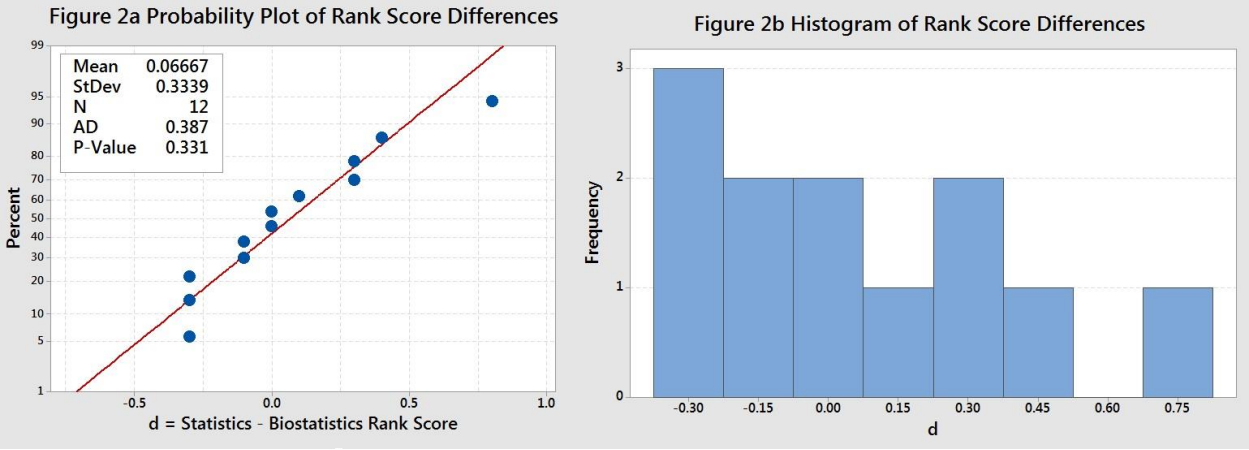
The resulting list of 69 schools however presented a problem for our analysis. Twelve schools have two programs listed (one for statistics and one for biostatistics) as shown in Table 4. Duplicate programs pose a problem in the construction of the regression models with the non-program specific metrics, since these are the same for both programs at the same school. Therefore, if two programs at the same institution have different rank scores, but are fit with the same set of covariates, this will most likely reduce the fit of the model artificially. Removing all the biostatistics programs is not an option (since eliminating the 24 biostatistics programs from the 69 programs would amount to eliminating about 35% of the data set). From Table 4 however, we note that the scores from both programs for each school are very similar: the sum of the differences is 0.8 and the mean difference is about 0.067. This observation suggests a hypothesis test on the mean of paired differences from a sample be used to determine if replacing each pair of scores with their mean is reasonable. Therefore letting d_i = (Statistics rank score for school i) – (Biostatistics rank score for school i), for i =1,…,12, the paired t-test

$$H_0 : \mu_d = 0$$
$$H_1 : \mu_d \neq 0$$

has a p-value of 0.5, indicating there was not sufficient evidence to conclude the mean of the differences was not equal to zero.

Table 4: Schools with Two Ranked Programs				
Number	**School**	**Stats Score**	**Biostats Score**	**d = Stats-Biostats Scores**
1	Berkeley	4.7	3.9	0.8
2	Harvard	4.3	4.6	-0.3
3	Washington	4.3	4.6	-0.3
4	UNC	3.7	4	-0.3
5	U of Mich	3.9	4	-0.1
6	Columbia	3.7	3.4	0.3
7	U of Minn	3.7	3.6	0.1
8	UCLA	3.4	3.4	0
9	U of Iowa	3.3	3	0.3
10	Yale	3.3	3.3	0
11	U of Pitts	2.8	2.9	-0.1
12	U of S Car	2.4	2	0.4

Since our sample size is relatively small, it was necessary to check if the distribution of the differences was normal (or at least symmetric). The p-value for the Anderson-Darling test for normality and the normal probability plot of the differences did not show evidence that the differences were not normal, however, the histogram of the differences did not look very symmetric (see Figure 2).



Since the t-test may not be an appropriate way to analyze this data set, a non-parametric test, the Wilcoxon signed rank test, was employed. The p-value of the test statistic (computed in R) was found to be 0.84, indicating that there was not sufficient evidence to conclude that the median of the two groups of scores are different. Based on these results, the two scores for each of these 12 schools were replaced with their average, leaving us with 63 data values.

With the data cleaning and preparation completed, the regression analysis was performed. Most of the regression analysis was performed using Minitab 17 Statistical Software (2010). The basic approach taken was to use simple linear regression on each individual continuous covariate. For each simple linear regression, the estimated regression coefficients, ANOVA table, summary statistics, and model significance were recorded along with the "four in one" plots of the residuals to check if the model assumptions were met. A model for a covariate whose values spanned two orders of magnitudes was also run using the square root and log transformations, and the form of the covariate with the largest R^2 value was chosen. Outliers were identified as those data values whose standardized residual had a magnitude exceeding three. For all such values, a further analysis was done to verify that the data value was valid and should be included in the data analysis. Although influential data values were recorded in Minitab, no further analysis with them was performed. After each covariate was appropriately transformed and inspected, its scatterplot with the response variable was examined. If the scatterplot appeared to be best fit by two lines, the indicator variable "control of institution" was used to see if it significantly improved the model. Finally, some of the metrics were so uniform (such as the admission factors) indicating no use in modeling, they were not used any further in the analysis. Both backwards elimination and forwards addition stepwise regression were performed in Minitab with all the covariates from each class of data types (program and non-program specific metrics). If the two methods yielded the same

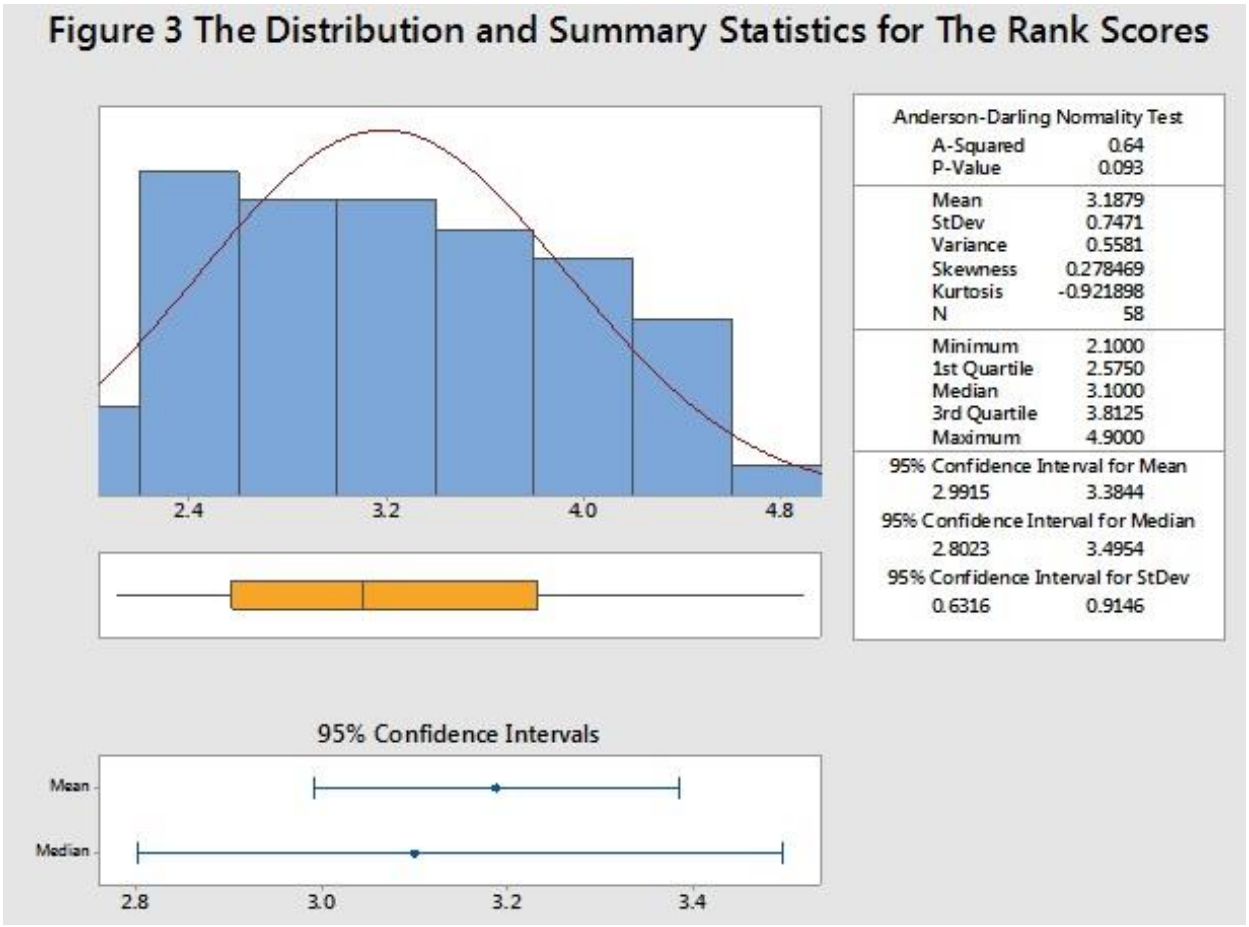
model, it was chosen as the best overall model. If the models from the two methods were not the same, then the best model was chosen (which was defined as the model with the smallest standard error, largest adjusted R^2 value and a C_p Mallows value that was not much greater than the number of predictors).

SECTION 3 RESULTS

3.1 Exploratory Data Analysis for Response Variable, Rank Scores

Exploratory analysis was performed on the rank scores and the results are shown in Figure 3.

Figure 3 The Distribution and Summary Statistics for The Rank Scores



Anderson-Darling Normality Test	
A-Squared	0.64
P-Value	0.093
Mean	3.1879
StDev	0.7471
Variance	0.5581
Skewness	0.278469
Kurtosis	-0.921898
N	58
Minimum	2.1000
1st Quartile	2.5750
Median	3.1000
3rd Quartile	3.8125
Maximum	4.9000
95% Confidence Interval for Mean	
2.9915	3.3844
95% Confidence Interval for Median	
2.8023	3.4954
95% Confidence Interval for StDev	
0.6316	0.9146

The optimal bin size for the histogram using the Diaconis-Freedman rule gave only five bins, so a bin size of 0.4 (roughly half a standard deviation) was chosen. The distribution appears to have a slight right skew, consistent with the mean (3.1879) being slightly greater than the median

(3.1). It is important to note that the rank scores have a range of only 3 and contains a large number of values (originally 87 and after data cleaning, 63).

3.2 Model Building for Program Specific and Non-Specific Metrics

Each single covariate was analyzed separately. This was done as a form of EDA to determine if a transformation was required, interaction terms with categorical variables were needed, or the model assumptions were violated. This section gives an overview of the analysis for each covariate by considering first the categorical and then the continuous covariates.

a) Categorical Covariates

There were seven possible categorical covariates. All were obtained from IPEDS[5]. One of the variables indicated if a school was a private or public institution (called "control of institution"), and it was included because it had been shown in a previous study to be a significant variable in modeling graduate program rankings (Grunig, 1997). The remaining six covariates were a group of secondary school requirements necessary for an applicant to be considered for admission to a school. Table 5 below shows the six covariates in this class along with the coding scheme for each.

[5] IPEDS (which stands for Integrated Postsecondary Education Data System) is a branch of the U.S. Department of Education's National Center for Education Statistics (NCES) that gathers information from every college, university, and technical and vocational institution that participates in the federal student financial aid programs.

Table 5: Metrics For Admissions Policy			
Number	**Metric**	**Coding in data file**	
1	Open admission policy	No Yes	0 1
2 **3**	Secondary school GPA Secondary school record	None Recommended Required	1 2 3
4 **5** **6**	Secondary school rank Completion of college-preparatory program Recommendations	Don't know None Recommended Required	0 1 2 3

An analysis of the admissions based criterion showed that none of the factors seemed suitable for the regression model. For each of the six factors considered, all the schools had the same level (such as open admissions and academic records), the distribution of rank scores for the different levels of the factors were all similar (such as school GPA and secondary school rank) or the counts of some of the levels were too small (such as completion of college preparatory programs and recommendations). In addition, since none of the scatterplots between the rank scores and the single continuous covariates suggested the need for the introduction of a categorical variable (i.e. they did not seem to follow two or three different distinct lines) the only categorical covariate utilized in the regression analysis was the indicator variable control of institution.

b) Non-Program Specific Metrics Continuous Covariates

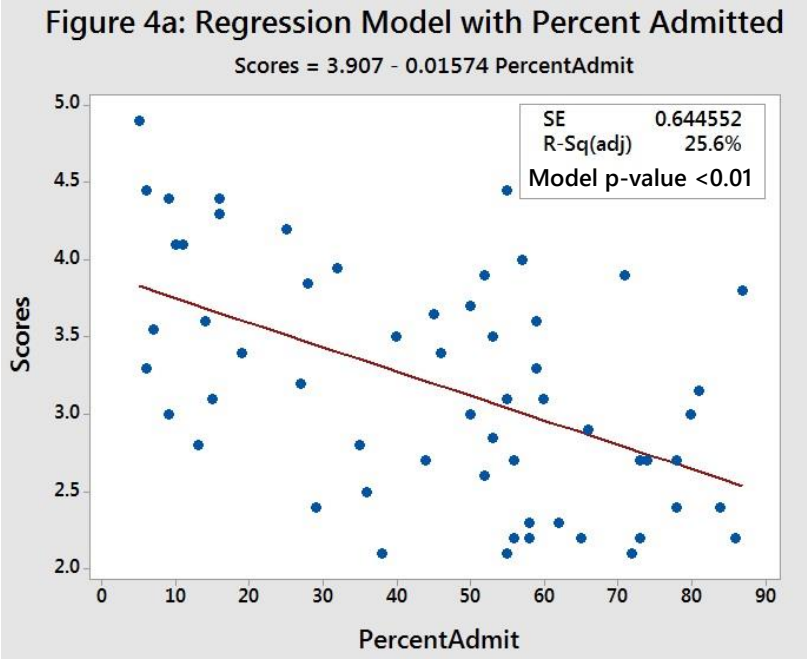
Eight continuous covariates are non-program specific metrics (see Table 3). The goal is to construct the best regression model with all of these nine covariates, the eight continuous and one categorical.

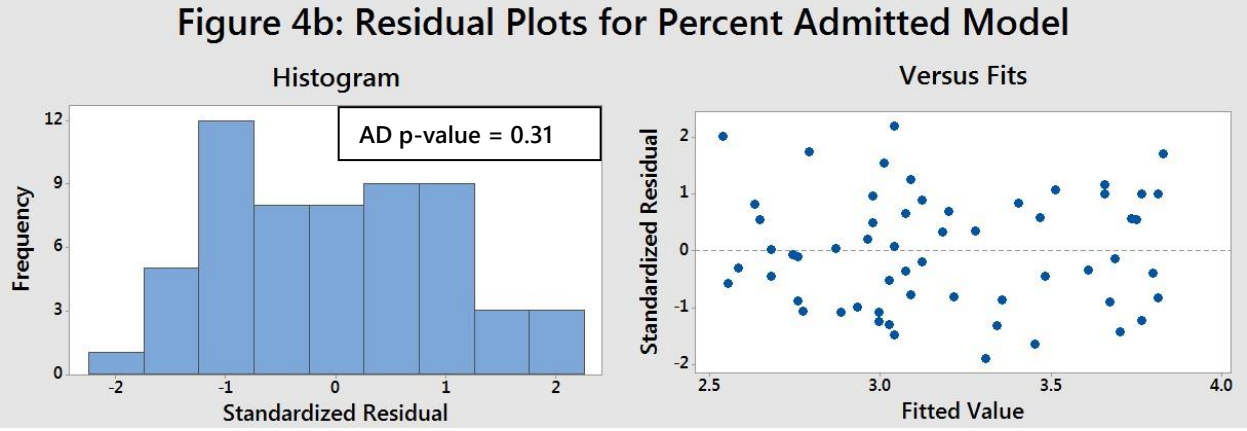
The first step was to perform simple linear regression on each individual covariate. The resulting model was inspected to check if the covariate was significant, a variable transformation was not necessary, and the model assumptions were not violated. For each covariate, two plots will be examined. The first plot (labeled with the letter "*a*") shows the scatterplot, the estimated regression line and its equation, the adjusted R^2 and the standard error. The second plot (labeled with the letter "*b*") shows a histogram of the residuals along with a scatterplot of the residuals against the fitted values and the p-value of the test statistic for the Anderson Darling (AD) test for normality.

The first six metrics (percentage students admitted, admissions yield, and the first quartile and third quartile for math SAT scores and critical reading SAT scores) were chosen due to the studies which have indicated these metrics are related to admissions selectivity and are significantly correlated with graduate program rankings obtained from reputational surveys. The remaining two, endowment and web presence, were included because they are metrics used in weighted ARS.

Percent Admitted:

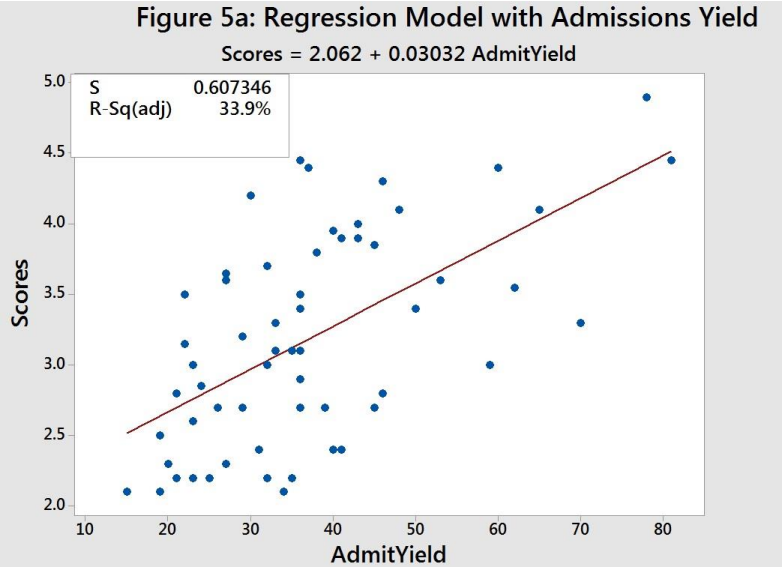
The percentage of students who applied to a school and were accepted was obtained from IPEDS. Figure 4a shows no variable transformation appears necessary and the covariate is significant. No influential points were detected. Two outliers both having standardized residuals around 2.2 were observed and not considered worth further investigation. Figure 4b suggests that the model assumptions are not violated.

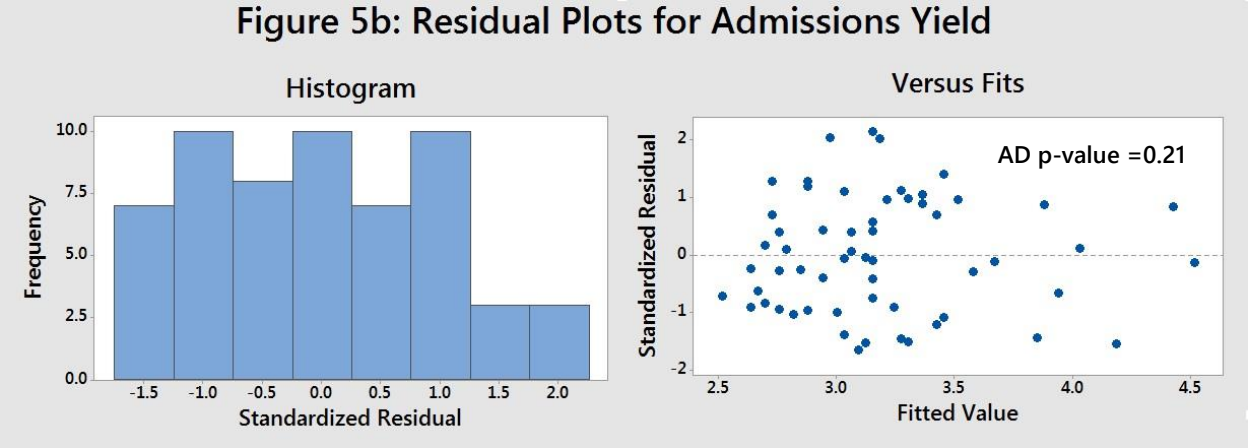




Admissions Yield:

The admissions yield is the percentage of admitted students that chose to attend. Figure 5a shows no variable transformation is necessary and the covariate is significant. Three outliers were detected. Since their standardized residuals were all below 3, no further investigation was performed. Figure 5b suggests the model assumptions are not violated.





SAT Scores:

IPEDS provides the first and third quartile scores for the math, critical reading and writing portions of the SAT exam. Because the writing scores were not provided by many schools on the list, only the critical reading and math scores were used. SAT scores are usually considered important general metrics for the quality of an academic institution. Also, analyses of reputational survey based graduate program rankings have shown there is a significant correlation between a school's composite SAT scores and its program's rank score (Fairweather, 1988 and Grunig, 1997). Rather than use the composite score, the scores for the individual parts of the exam were used since the math scores might be high for most of the institutions, but the critical reading scores might better differentiate the top institutions from the rest.

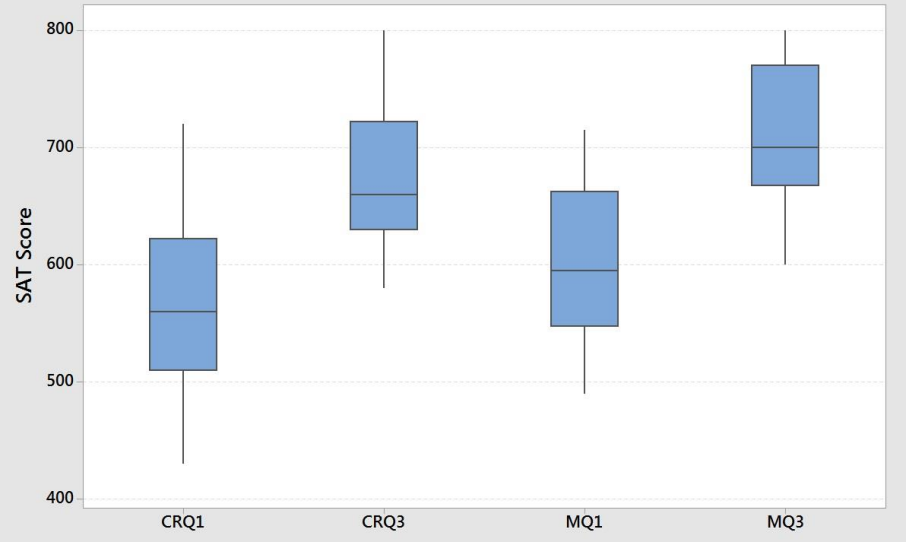
A major problem with using the SAT scores in regression models is they are highly correlated. Table 6 below shows the Pearson correlation coefficients for the various pairs of SAT scores (where "CR"/"M" indicate scores from the "critical reading"/"math" portions of the exam and Q1/3 indicate the values for the first and third quartiles).

Table 6: SAT Score Correlations

	CRQ1	CRQ3	MQ1
CRQ3	0.967		
	0.000		
MQ1	0.934	0.931	
	0.000	0.000	
MQ3	0.848	0.908	0.956
	0.000	0.000	0.000

One way to deal with a number of significant, yet highly correlated covariates is to use Principal Component Analysis (Jolliffe, 2002). It is particularly well suited to the covariates in this problem, since, as Figure 6 shows, they are all measured on the same scale and have roughly the same variance.

Figure 6: Distribution of SAT Scores



The MINTAB output from the PCA, given in Table 7, shows there is one dominant eigenvector, which accounts for roughly 94% of the total variation in the variance-covariance matrix. It is a linear combination that weights the scores about equally.

Table 7: PCA Output for SAT Scores

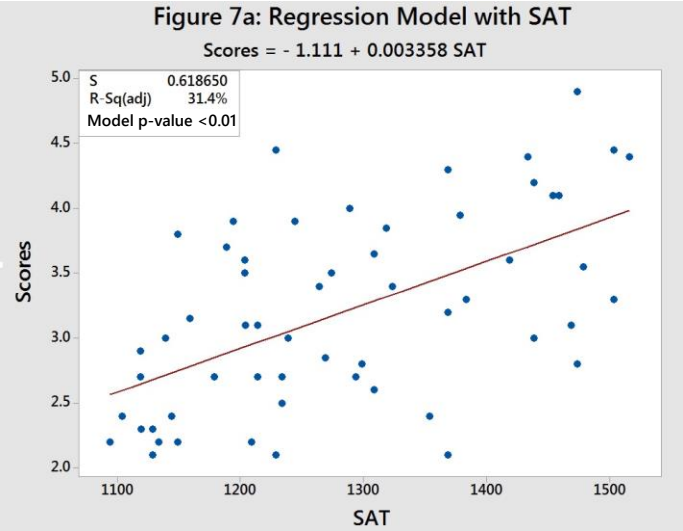
Eigenvalue	3.7718	0.1702	0.0527	0.0052
Proportion	0.943	0.043	0.013	0.001
Cumulative	0.943	0.986	0.999	1.000

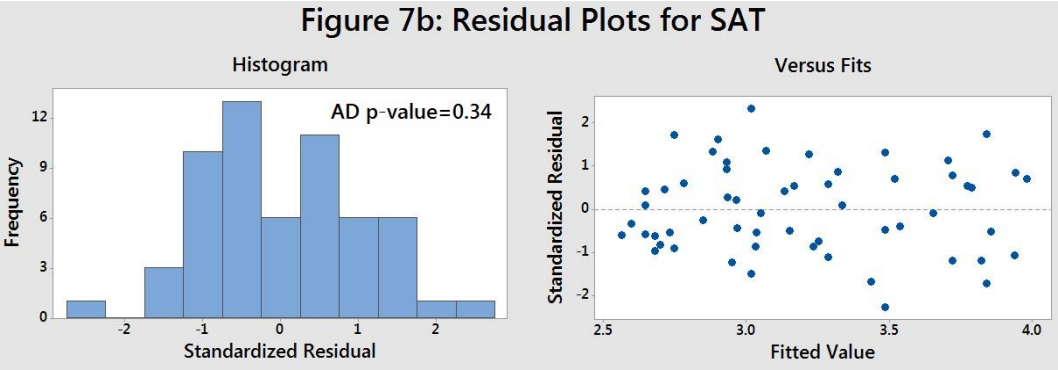
Variable	PC1	PC2	PC3	PC4
CRQ1	0.497	-0.608	-0.269	0.557
CRQ3	0.504	-0.307	0.659	-0.467
MQ1	0.506	0.226	-0.650	-0.519
MQ3	0.492	0.696	0.266	0.450

From the PCA, a new covariate, SAT, is defined as the eigenvector corresponding to the largest eigenvalue, computed in MINITAB using the weights in Table 7:

$$SAT = 0.497 * CRQ1 + 0.504 * CRQ3 + 0.506 * MQ1 + 0.492 * MQ3 \qquad (5)$$

The estimated regression line with the covariate SAT is shown in Figure 7a. The output indicated no influential points. Two outliers were observed with standardized residuals magnitudes less than 3. The residual analysis, from Figure 7b, indicates that the model assumptions are satisfied.

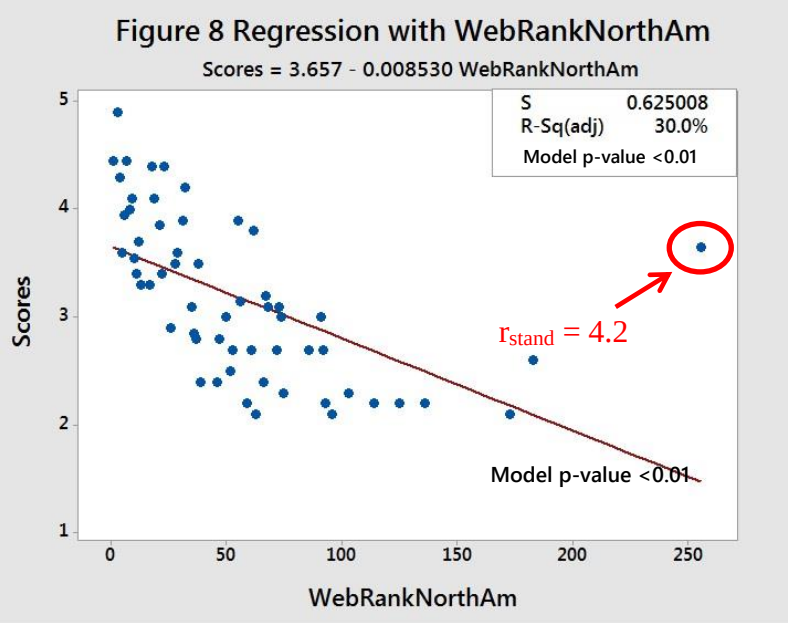


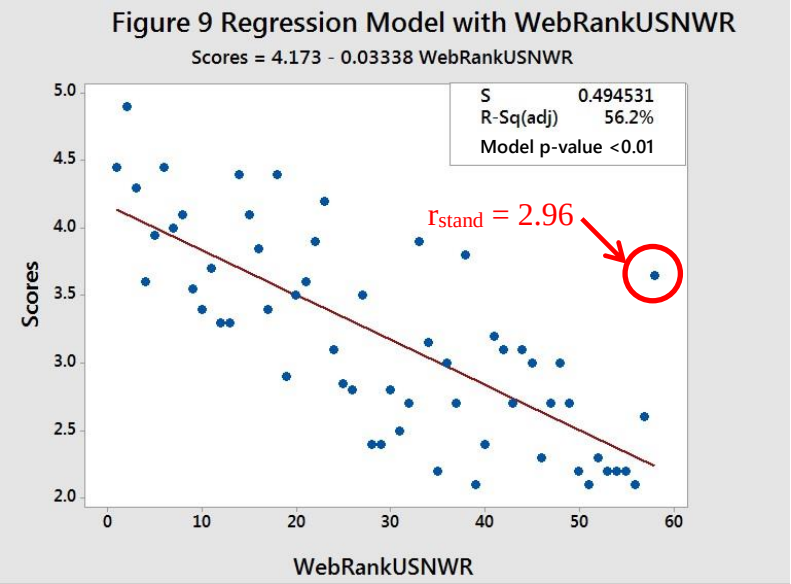


Web Presence

The internet is widely recognized as the "information superhighway" and has come to play a major role in education and research. Web presence is an attempt to quantify how much an institution meaningfully contributes to the online production and dissemination of information. Although "web presence" is difficult to define, the most widely used method is called "webometrics" produced by Cybermetrics, a research group from the Spanish National Council on Research. Webometrics publishes a yearly comprehensive ranking of universities worldwide based on their web presence on the website (Webometrics, 2016). The web presence computed by webometrics is a measure of how many meaningful links an institution has on the internet (somewhat like a "high impact citation count" for the internet), and uses a methodology called "link analysis" (Aguillo, et. al., 2006). The data used in this analysis was from the 2015 rankings of all North American universities and the ranking for each school was obtained individually from the list. Every school in our analysis had a Webometric

ranking. However, since there were often large gaps in the rankings of universities (especially near the bottom of the web rankings), the webometrics rankings of the schools were ordered. Then their relative web rank was also used (so for example, the University of Florida was rated 22nd best among North American schools, but was the 17th best among the schools in our analysis). To distinguish between the two ways used to rank the schools, the variables are labeled "WebRankNorthAm" and "WebRankUSNWR". Figure 8 and 9 below shows the simple linear regression models for the North American and USNWR web rankings, respectively.





The University of Minnesota (Twin Cities) is an outlier in both models as indicated by the red circled data value and its standardized residual. This data value is an anomaly. Looking at the list of North American universities, it ranks 256, below the University of Minnesota, Duluth. This incorrect estimation is because the University of Minnesota, Twin Cities has several domain names which splits up its web presence and according to the director of the Webometrics research team, cannot be remedied[6]. To better estimate the University of Minnesota's web presence, its value was imputed from the other covariates (for the non-program specific metrics) in this study[7]. The logarithm[8] of the endowment had the largest correlation with web presence for both

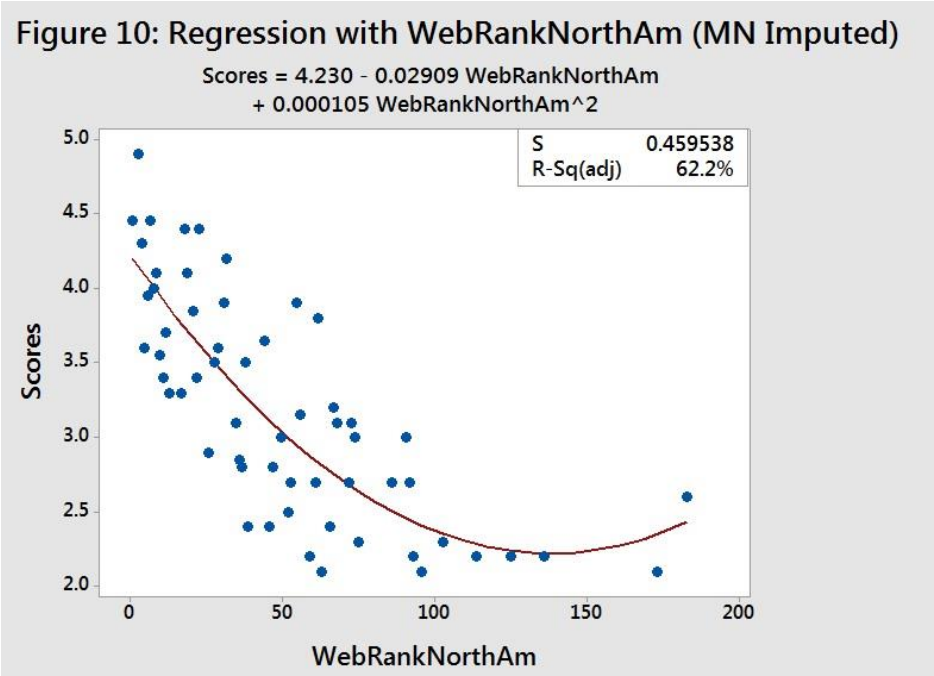
[6] Isidro F. Aguillo, personal communication, May 19th, 2006
[7] The imputation for the University of Minnesota was done post hoc the analysis for all the non-program specific metrics; hence the covariate "Endowment", which is considered in the next section, was used in the imputation process.
[8] All logarithms used are base 10

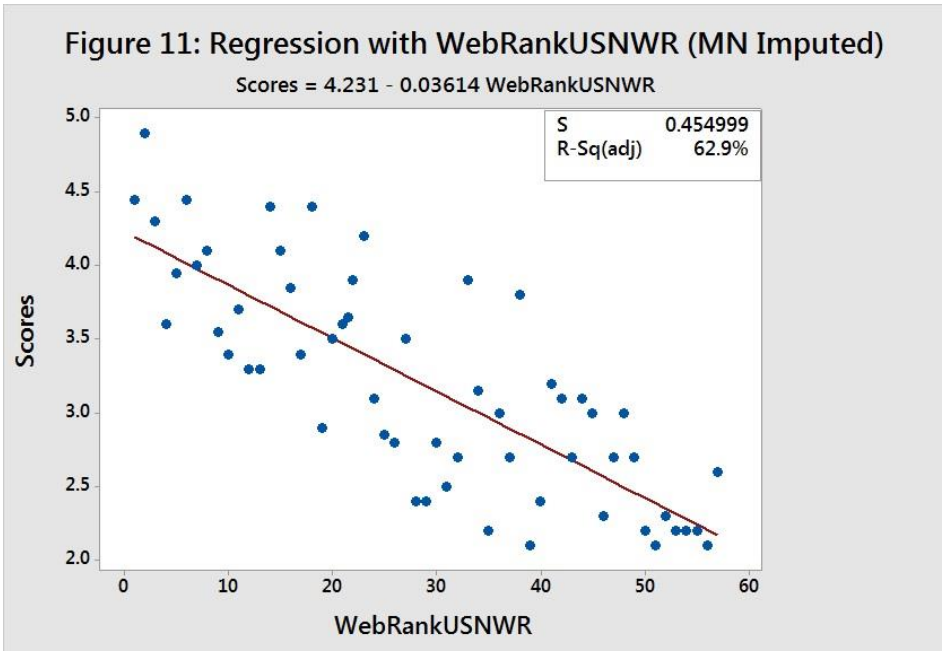
webometric rankings. Performing regression on web presence using the endowment, the following regression equations were obtained:

$$WebRank\,USNWR = 157.4 - 20.84 * Log(Endow) \quad (6.a)$$
$$WebRank\,NorthAm = 338 - 45.03 * Log(Endow) \quad (6.b)$$

Substituting the Log(Endow) value for the University of Minnesota of 6.5182, into equations (6.a) and (6.b), the imputed values of WebRankUSNWR = 21.56 and WebRankNorthAm = 44.49 were obtained. With the imputed web ranking values for the University of Minnesota, the regression model for both covariates improved significantly.

As Figure 10 shows, the North American rankings are fit well by a quadratic model with the adjusted R^2 increasing from 30% to over 62%. The fit for the USNWR rankings also improves with its adjusted R^2 increasing from about 56% to about 63% as shown in Figure 11.

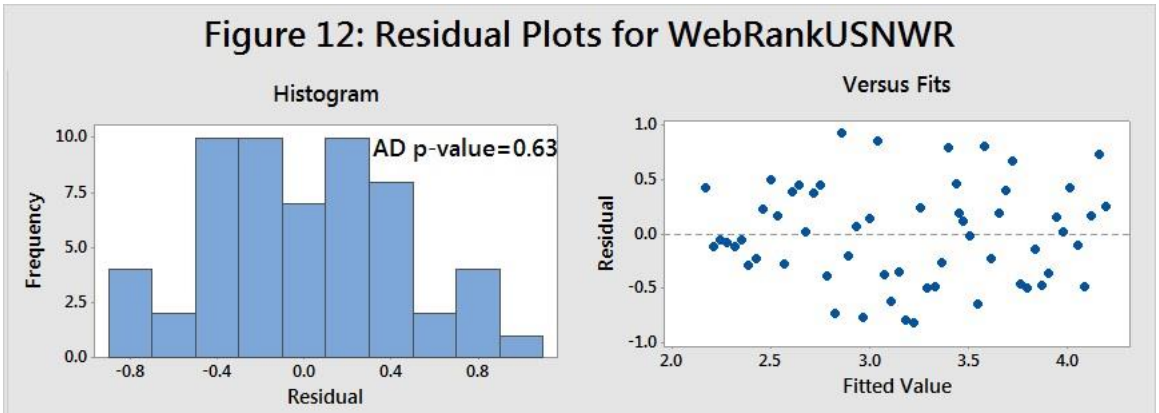




The final model summary statistics for both web rankings using the imputed value for the University of Minnesota are shown below in Table 8.

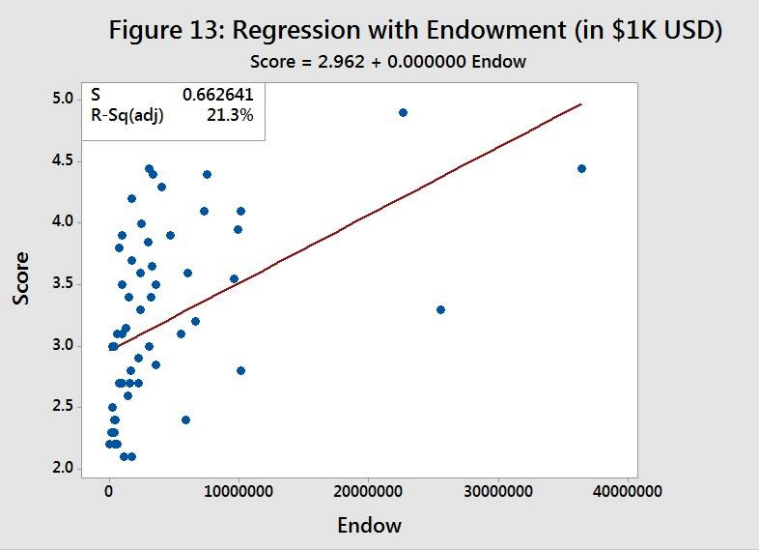
Table 8: Best Linear Regression Models for Two Representations of Web Rankings		
Covariate	**USNWR**	**North American**
Est Reg Eqn	Score = 4.231-0.03614 WebRank	Score = 4.23 - 0.0291 WebRank + 0.00011 WebRank^2
S	0.455	0.456
Adj R^2	62.9%	62.2
AD p-value	0.63	0.57

Since the goodness of fit was about the same for both covariates, the more parsimonious model using the USNWR ranked covariate was used. The residual plot shown in Figure 12 indicates that the model assumptions are not violated.



Endowment:

Past research has shown a significant "prestige" measure that correlates well with the USNWR academic rankings is alumni giving. Since finding reliable data on this quantity was difficult, "endowment" was used as a proxy for alumni giving. However, obtaining reliable estimates for an institution's endowment is also not without its challenges. There are different accounting systems by which an institution's endowment may be computed and the endowment for a specific branch of a state university system can be difficult to obtain. The approach taken in this study was to use the estimates given in IPEDS for 2014 endowment assets. For each school, its IPEDS endowment estimate was compared to the estimated endowment on the Wikipedia page for the school. When the two values differed by more than $10 million, further research was conducted on the school's website to resolve the discrepancy. Although discrepancies for three schools were found, all were resolved. All endowment values were in USD and expressed in units of $1K. Simple linear regression run with endowment, shown in Figure 13, suggests that a transformation might be appropriate.

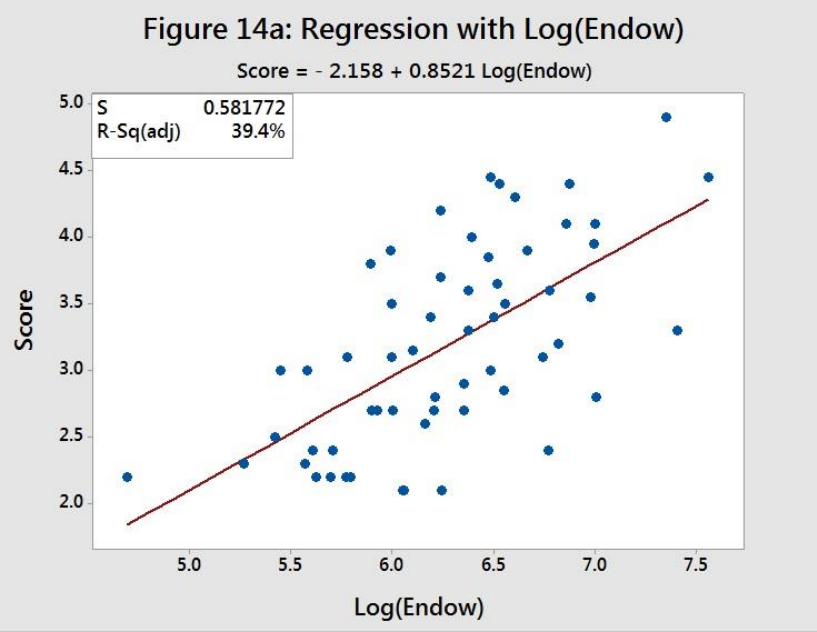


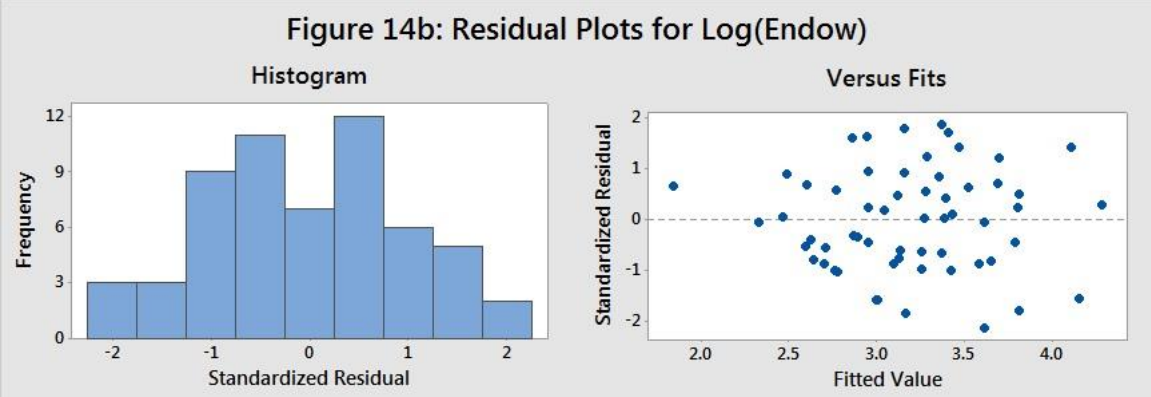
Both the square root and log transformations were performed. The results shown below in Table 9 suggest the log transformation appears best.

Table 9: Model Summary Statistics for Different Transformations of Endowment			
	Endow	**Sqrt(Endow)**	**Log(Endow)**
S	0.663	0.613	0.582
Adj R^2	21.3%	32.63%	39.35%
Model p-value	<0.01	<0.01	<0.01
AD p-value	0.09	0.146	0.591

The scatterplot and residual plots, in Figures 14a and 14b respectively for the log transformed covariate support the summary statistics that it produces the best model that does not violate the model assumptions. One outlier was flagged, but its standardized residual was about 2.2, and therefore not considered to be worth further investigation. Two influential values were also found (Harvard and SUNY Albany). When these two data values were removed, the simple linear regression showed little change in the model parameters and summary statistics. Interestingly, the one influential data value in the new model (when Harvard and Albany were dropped) was Yale. Harvard and Yale are first and second in endowment. As one can see from the scatterplot, these few schools with very large endowments (even after the log transformation) are far enough away from the other covariate values that they still are

influential. It was determined nothing further could be done about these influential data values and since they did not seem to significantly affect the model overall, they were ignored.





Best Multiple Linear Regression Model

The preceding analysis of the single continuous covariates leaves five variables with which to build the best regression model. Although there are many methods for

building regression models, the approach taken in this analysis is to perform stepwise regression using both forward addition and backwards elimination. If both methods yield the same model, then there is confidence that the final model is probably the best model overall.

Stepwise regression using all 5 covariates was run in MINTAB with an α to enter value of 0.25 and with an α to remove value of 0.1. Both methods produced the same model; the summary statistics are shown below in Table 10.

Table 10: Best Regression Models with Non-Program Specific Metrics

Summary Statistic	Value
S	0.443
Adj R^2	64.87
C_p Mallows	1.6

The summary statistics for the coefficients of the final estimated regression equation

$$Rank\ Score = 2.685 + 0.001099\ *\ SAT - 0.03137\ *\ (WebRankUSNWR) \qquad (7)$$

are shown in Table 11:

Table 11: Summary Statistics Related to Estimated Model Coefficients

Term	Coeff	SE	Coeff T-Value	p-Value	VIF
Constant	2.685	0.786	3.42	0.001	
SAT	0.001099	0.000554	1.98	0.052	1.44
WebRankUSNWR	-0.03137	0.00426	-7.37	0.000	1.44

Looking at the regression coefficients, we see that the coefficient for web rank is negative; however, this is reasonable since the rankings are from best (1) to worst (58).

c) Program Specific Metrics Continuous Covariates

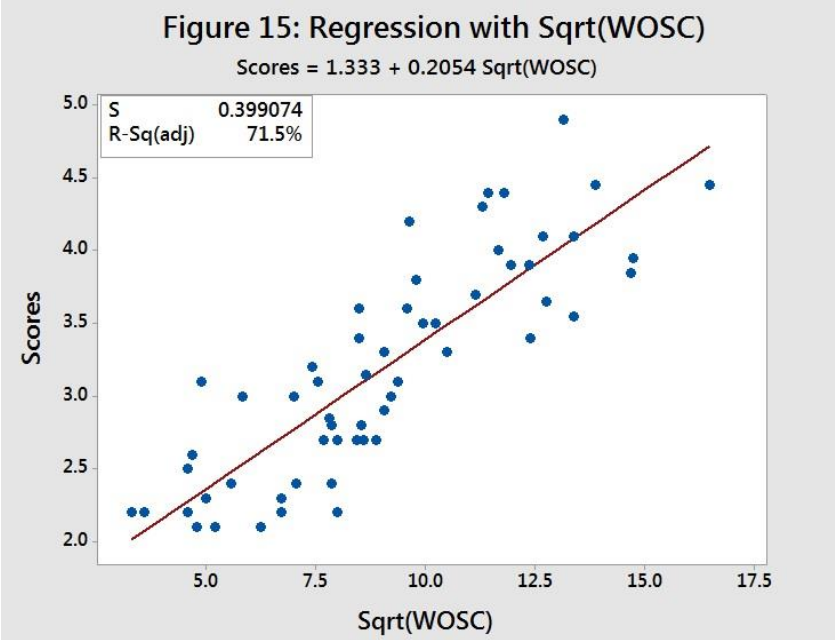
There are four continuous covariates that are program specific metrics (see Table 3) and the indicator variable, control of institution. The goal is to construct the best regression model with all of these five covariates.

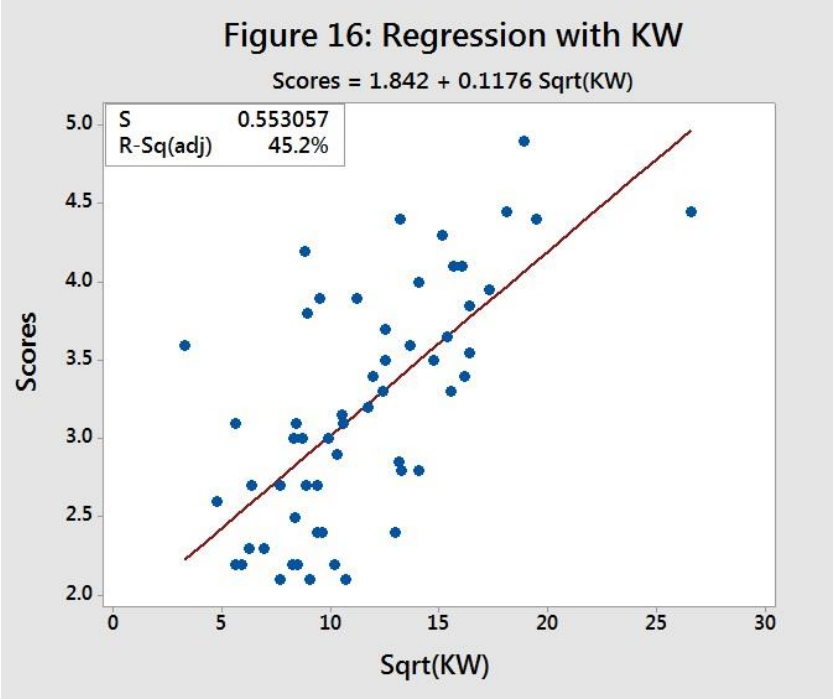
The first step was to perform simple linear regression on each individual covariate. The resulting model was inspected to check if the covariate was significant, variable transformations were not necessary and the model assumptions were satisfied. For each covariate, two plots are displayed. The first plot (which will be labeled with the letter "*a*") gives the scatterplot, the estimated regression line and its equation, along with the p-value for the model, adjusted R^2 and standard error. The second plot (labeled with the letter "*b*") shows a histogram of the residuals along with a scatterplot of the residuals against the fitted values and the p-value of the test statistic for the Anderson Darling test for normality.

Number of Publications:

While the number of publications is an important metric, as it is a measure of research productivity, it is difficult to define precisely. Since a large number of publications alone is not evidence of quality research, some analysts have chosen to measure publications using citation counts, and more complicated metrics (like the h-factor that count citations in high impact journals). Many of these metrics are simply beyond the scope of this study since they must be done for each individual researcher at each institution. Thus, only metrics which looked at the publications of each department *in toto* were used.

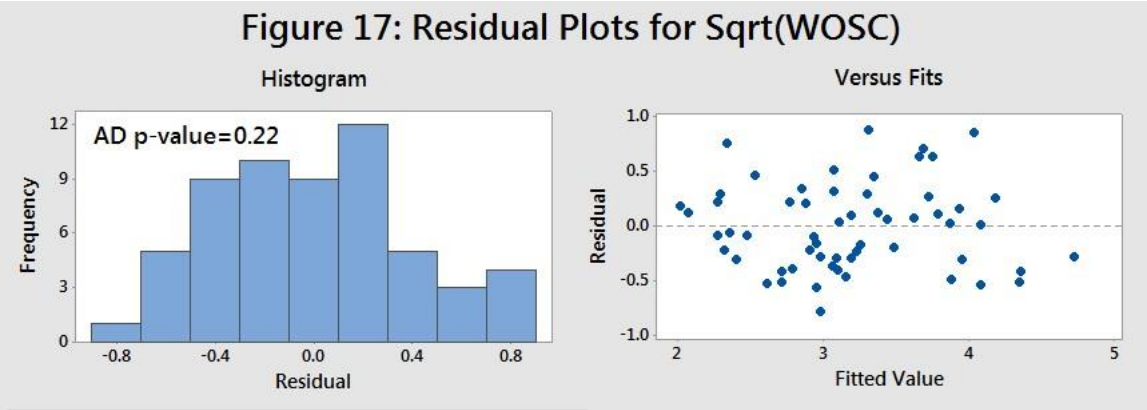
The data source used to gather information about the number of publications for each program was Thomas Rueters Web of Science, accessible through the library system at Ball State University. Each institution was searched using the "enhanced search option" which insures that all institutions and schools are included in two ways. Each method counted the number of articles published in peer-reviewed journals in 2014. The first search counted the number of publications by each institution using the Web of Science keyword subject search "probability and statistics". The counts from the first search are labeled "WOSC" (World of Science Category). Due to concerns that this keyword search might not adequately count articles in biostatistics, a second method for searching the publications was chosen using the keyword search "statistics" (which found all articles whose keywords contained the string "statistics"). The counts from this search are labeled "KW" (for keyword). The scatterplots and summary statistics for each publication metric (after they are transformed appropriately) are shown below in Figures 15 and 16.





Figures 15 and 16 clearly show that the square root of the WOSC counts is the best choice of covariates for modeling the USNWR rank scores. Two outliers were detected, although both had magnitudes of less than 2.5 and were not investigated further. The one influential data value, Harvard University, was due to the fact its value is so much

larger than the other schools. It lies at the extreme end of the covariate scale even after the square root transformation. It was determined nothing could be done about this data value and even though its leverage is twice the threshold value of 3/n. When the data value was removed and regression run, neither the adjusted R^2 nor model coefficients change significantly; hence it was ignored. Below in Figure 17 are the residuals for this model, which indicate that the model assumptions are not violated.

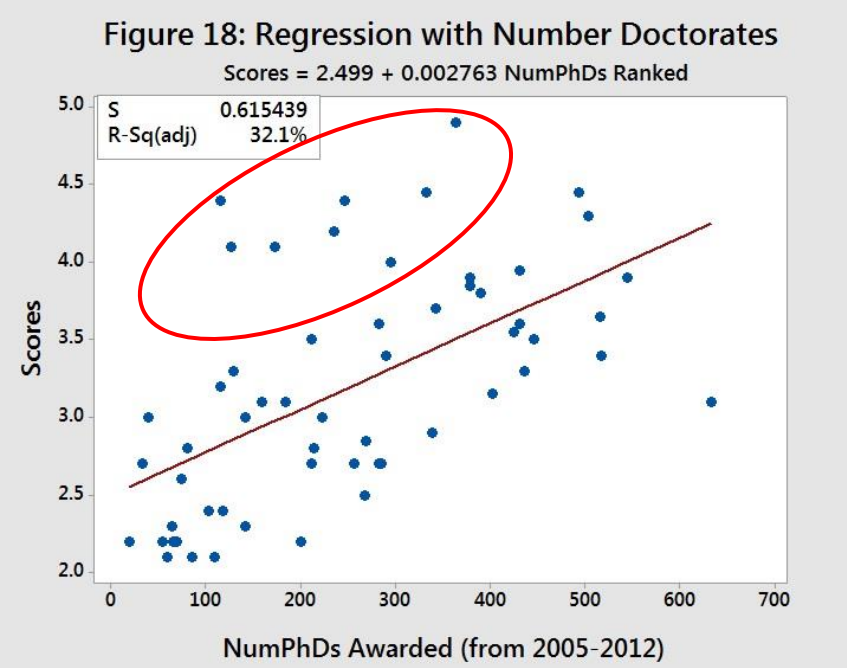


Number of PhDs:

The number of doctorates awarded by each program over 2005 to 2012, the years over which the survey spanned, was used as the metric. The standard source for this data is the NSF[9]. To insure that both biostatistics and statistics doctorates were included, a combined search was performed using the specific NSF code for "biostatistics" and "general statistics". Regression analysis shows that there are three minor outliers (with residuals having magnitude less than 2.6) and one influential data value. A specific concern raised by Figure 18 is the group of data values that are clustered high above the regression line (circled in red). An examination of these

[9] National Science Foundation

schools shows they are all private and hence raises the possibility of needing to introduce the categorical variable "control of institution."



Therefore, this categorical variable (which is an indicator for whether an institution is private or public) was introduced into the regression model. The best model yielded two groups with different y-intercepts. The adjusted R^2 increased from 32% to 46.94% and the standard error decreased from 0.615 to 0.544. The p-value for the significance of the coefficient for the indicator variable is less than 0.001. The scatterplot of the regression lines to fit these two groups, shown in Figure 19a, indicates that institutional control is a significant covariate. Figure 19b suggests that none of the model assumptions are violated.

Figure 19a: Regression with Institutional Control Added

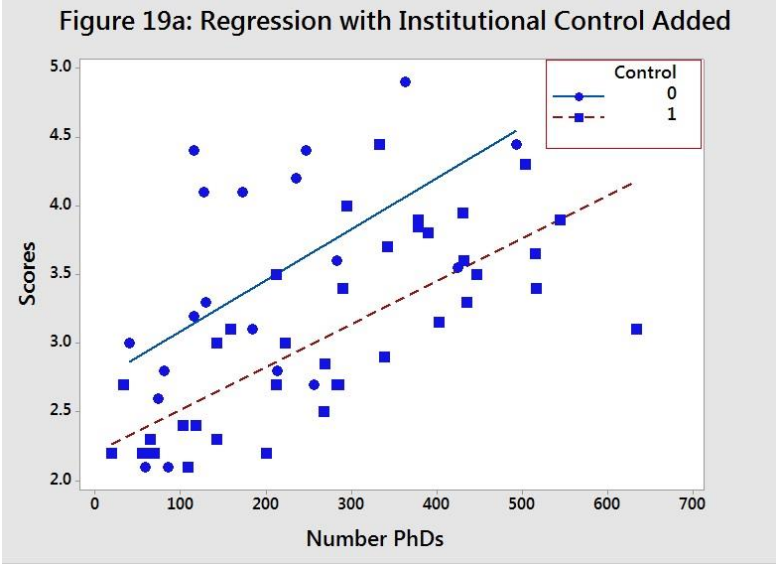
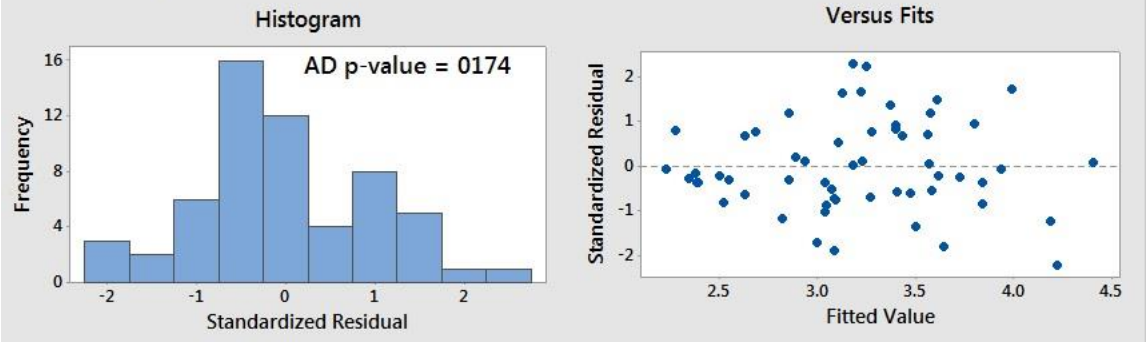


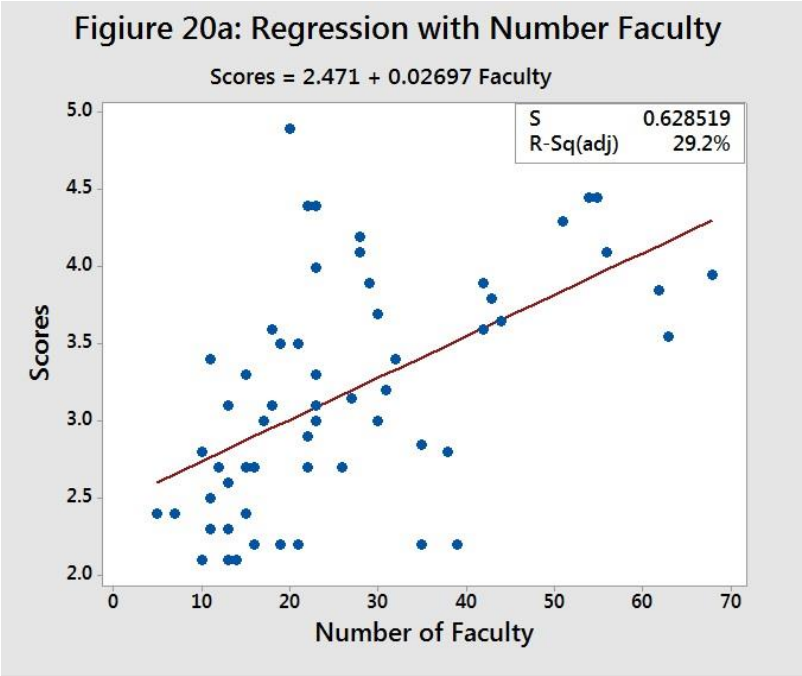
Figure 19b: Residual Plots for Number PhDs + Control

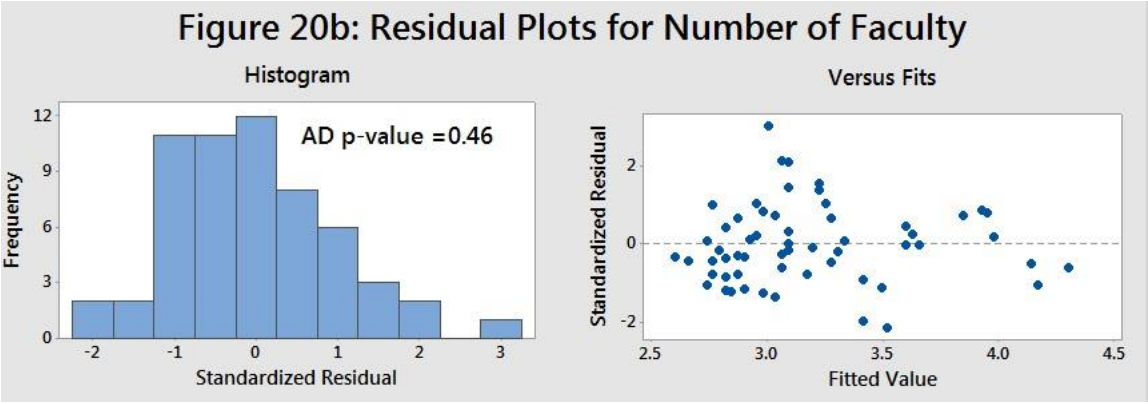


Number of Faculty:

Previous studies of graduate program rankings indicate that department size affects reputational survey results (Fairweather, 1998). The metric of department size chosen was the number of potential research faculty. These counts were obtained by going to website for the departments of each ranked school and counting the number of faculty, then eliminating those who were obviously not potential faculty to advise a doctoral student (such as adjuncts, instructors, and teaching faculty). As both the regression

statistics and the scatterplot in Figure 20a indicate, the model fit is not very good owing to some extreme outliers (Stanford University in particular). Introduction of the indicator variable "institutional control" improved the model fit marginally. Despite the outliers, the p-value for Anderson-Darling test statistic of 0.46 indicating that perhaps the model assumptions were not violated (although neither the histogram or fitted plots for the standardized residuals visually seem very good).

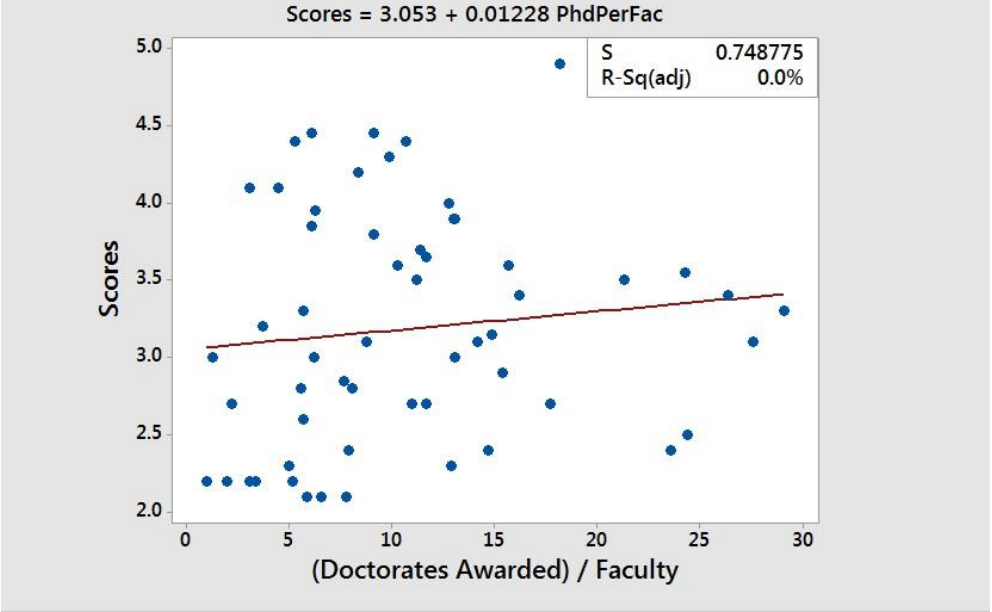




Doctorates Awarded to Faculty Ratio:

The last metric considered was the ratio of doctorates awarded to faculty members. This ratio was obtained directly from the last two metrics. As Figure 21 shows, this covariate surprisingly has no significant predictive power and is therefore dropped from our modeling.

Figure 21: Regression with Doctorates Awarded : Faculty Ratio



Best Multiple Linear Regression Model

The best regression model using stepwise regression techniques with the three continuous covariates, Sqrt(WOSC), Number Doctorates Awarded, and Number Faculty, along with the indicator variable institutional control, was performed. Unlike the case of the non-program specific metrics, backwards and forwards elimination did not yield the same model. For this reason "best subsets" (giving best 2 models of size 1, 2, and 3) was also run to insure that the best model was not missed. The results from all three model building methods are shown in Table 12.

Table 12: Comparison of Models with Program Specific Metrics				
Method	**Model Variables**	**Adj R^2**	**SE**	**C_p Mallows**
Backwards	Sqrt(WOSC) Control	73.43%	0.38504	3.2
Forwards	Sqrt(WOSC) Number Doctorates Control	73.9%	0.38160	3.2
Best Subsets	Same model as Forwards			

Thus, the best model obtained uses Sqrt(WOSC), Number Doctorates and also control.

The final best regression model using the program specific metrics is given below:

$$USNWR\,Rank\,Score = 1.625 - 0.316\,Control + 0.00671 * (Number PhDs) + 0.1781\sqrt{WOSC} \quad (8)$$

Details of the regression output are shown below in Table 13.

Table 13: Summary Statistics Related to Estimated Model Coefficients					
Term	**Coeff**	**SE**	**Coeff T-Value**	**P-Value**	**VIF**
Constant	1.625	0.190	8.54	0.000	
Control	0.1781	0.0234	7.61	0.000	2.04
Number PhDs	0.000671	0.000475	1.41	0.163	2.15
Sqrt(WOSC)	-0.316	0.118	-2.67	0.010	1.22

d) Application of the Test for Validity

The statistical hypothesis test for the validity of the reputational survey based USNWR graduate statistics program rankings is to see if the rankings "fit better" with metrics that are specific to the statistics programs. As outlined in section 1, this can be expressed as the following hypothesis test on the adjusted R^2 values for the best fit regression models using the program specific (p) and non-program specific (~p) metrics as:

$$H_0 : \left(R^2_{adj}\right)_p \geq \left(R^2_{adj}\right)_{\sim p}$$
$$H_1 : \left(R^2_{adj}\right)_p < \left(R^2_{adj}\right)_{\sim p}$$

Although these two statistics do not come from the same distribution (hence we cannot construct a test statistic), we can compute the confidence interval for both adjusted R^2 values. If the confidence intervals overlap, that will be interpreted as equivalent to failing to reject the null hypothesis.

The 95% confidence interval for the adjusted R^2 value was computed using the function ci.R2 contained in the R package MBESS, based on a method which recursively evaluates the hypergeometric function for the density function of R^2 (Lee, 1971). The results for the two classes of metrics are shown below in Table 14:

Table 14: Confidence Intervals for Adjusted R^2

Metric	**95% Confidence interval for Adjusted R^2**
Non-Program Specific	$0.51099 \leq \text{Adj } R^2 \leq 0.78641$
Program Specific	$0.63174 \leq \text{Adj } R^2 \leq 0.84626$

Based on the overlap in the confidence intervals, one would fail to reject the null hypothesis and conclude there is not sufficient evidence to support reputational survey based USNWR statistics graduate program rankings are valid.

SECTION 4 DISCUSSION

4.1 Interpretation of Results

The goal of this thesis was to propose a statistical hypothesis test for the validity of reputational survey based graduate academic rating systems and to apply this method to the 2013 USNWR ranking of U.S. statistics doctoral programs. The underlying notion of the test came from previous studies that showed metrics related to the institution in which the program resided, but not the program itself, were often highly correlated with the graduate program rankings. For example, Fairweather (1988) showed that undergraduate selectivity was a significant covariate in a regression model for the National Academy of Science reputational rating of faculty (often used as a measure of a graduate program's quality). Other studies showed the USNWR rankings for different programs at the same institution had very high correlations (Webster, 1988) and they were often correlated to SAT scores (Grunig, 1997). While these observations suggest the possibility of a halo effect, none of the studies were able to test the validity of GPARS. Building a regression model for a ranking that contains a significant not program specific covariate is not evidence of the halo effect since only a limited number of covariates were studied. This leaves the possibility that a collection of program specific covariates may have produced a better model. Moreover high correlations between non-program specific metrics of academic quality and graduate program rankings are not surprising since top universities tend to be uniformly good in all areas of academics. The logic underlying the approach in this thesis is to realize if a GPARS is valid, then it must be modeled *best* by program specific metrics (of academic quality). Or to put it another way, we would have little confidence in a GPARS if we were able to find a model based entirely on non-program specific metrics which was superior to every model based entirely on program specific metrics. If we choose a parameter

like the adjusted R^2 value of a regression model as a criterion for judging which model is best, then the statistical hypothesis test is on the difference in the supremum of this parameter for the two collections of metrics. If enough regressors are chosen from both sets of metrics (i.e. program and non-program specific), we should be able to estimate the parameter well and construct a reliable hypothesis test.

This study applied the hypothesis test on the data over two time periods from reputational surveys taken by USNWR on 63 U.S. statistics doctoral programs. Based on previous studies, the number of publications in statistics, the number of statistics doctorates awarded, and the number of research faculty were chosen as program specific measures of quality of the statistic program. Metrics such as admissions selectivity, SAT scores, web presence, and endowment were chosen as metrics of the school's general quality that was not due directly to the quality of the statistics department. Data on all of these covariates were collected from reliable sources (almost exclusively from government educational databases) and the best fit models for the program and non-program specific metrics were found. A comparison of the adjusted R^2 values showed the program specific model (adjusted $R^2 = 73.9\%$) was better than the non-program specific model (adjusted $R^2 = 64.9\%$). Since the survey had a 39% response rate, these rankings were a sample from the population of all respondents and the R^2 values were themselves statistics. Since the distribution of the difference of the two adjusted R^2 values is very complicated, the hypothesis at a significance level α, using the program specific (p) and non-program specific (~p) metrics can be expressed as:

$$H_0 : \left(R^2_{adj}\right)_p \geq \left(R^2_{adj}\right)_{\sim p}$$
$$H_1 : \left(R^2_{adj}\right)_p < \left(R^2_{adj}\right)_{\sim p}$$

This is equivalent to the condition that the $(1-\alpha)100\%$ confidence interval of the adjusted R^2 value for the program specific metrics does not overlap with the confidence interval based on the non-

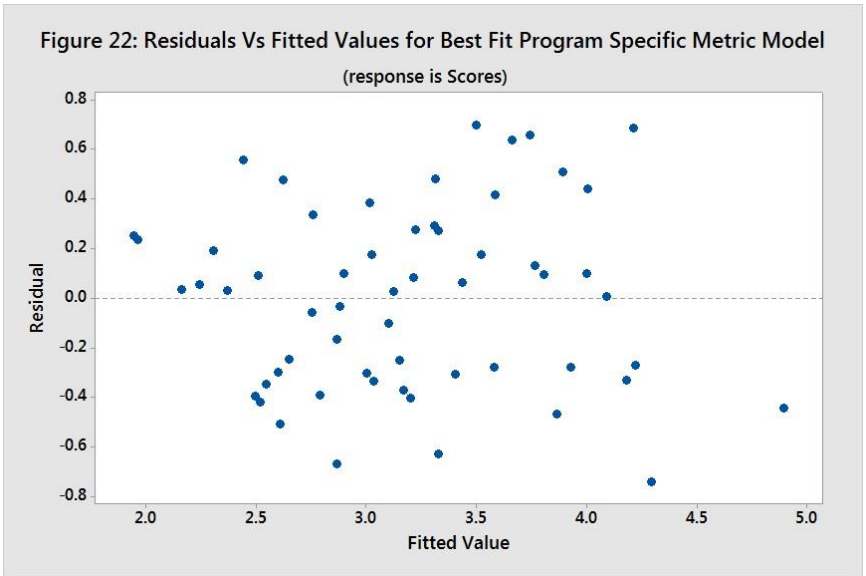
program specific metrics. When the decision rule is applied at α=0.05, the two confidence intervals overlap. From this result, it was decided that there was not sufficient evidence to conclude that the USNWR reputational survey based GPARS were valid for statistics doctoral programs.

4.2 Limitations of Study and Conclusions

There were several limitations of this study. A major area of concern in the analysis involves the regressors used in the modeling. First, only a selected number of regressors were chosen. The quality of any educational system is highly multivariate in nature, and so it is quite conceivable with a more complete list of regressors (for both the program and non-program specific metrics), the results of this hypothesis test would change. Second, the choice of metrics could also affect the results of this study. Apropos of this study, it is known from previous studies reputational surveys that purport to measure academic quality can be explained largely by program size and scholarly activity (King and Wolfe, 1987; Saunier, 1985). But there are myriad of different ways both of these quantities could be measured. In this study, program size was measured by the number of research faculty, although this may not have been the correct choice of metric. Large departments, for example, might also need large teaching staffs. Thus, the total number of individuals in a department who taught might have been a better measure of department size. It could also be that the number of doctoral students in the program, and not just the number of faculty, was an important measure of the program size. Scholarly activity, can also be measured many different ways (using total number of publication, number of books published, awards won, conferences attended, inter-institutional collaboration, and citation counts). Third, the quality of the data could also be an issue. Although the data set in this study came from the same database used by the ASA, there are substantial deviations between the two in some statistics (such as the number of statistics

doctorates awarded). Although it is perhaps natural to assume the ASA's data set is correct, there are clear errors in their data collection methods. The list of schools they provided is supposed to contain all doctoral statistics programs that have graduated five PhDs over the four year period 2009-2012. However, Bowling Green State University, which on their own website lists seven doctorates being awarded in this time period, was not included on the ASA list. Related to data quality, is a fourth issue of more specific and better defined datasets. For example, the methodology behind how World of Science counts the number of publications from each school in the category "probability and statistics" is not known, and so these estimates may under count publications involving interdisciplinary research.

Another limitation is that not all of the schools are equally sampled. The total number of surveys sent out was 348. At a response rate of about 39%, this means that roughly 136 surveys were collected. It is known from USNWR some programs barely received 20 rankings. Yet we can be relatively confident that some of the top schools (such as Harvard and Stanford) probably were ranked on at least 80% of the returned surveys (or about 109 surveys). This means around five times more surveys on which the top schools were ranked than the lower schools. Thus, the estimates for the higher rank scores are more accurate than those of the lower rank scores. Weighted least squares might be considered as one way to try to account for the accuracy of the scores, but as Figure 22 shows, there is no substantial difference in the residuals or their trend between the top ranked and lower ranked schools.



There are also potential problems with the way in which the rankings are computed. For example, the USNWR attempted to rank 87 schools in a range of 3 units. The resolution required for ranking these programs accurately is at most 3/87 = 0.034. However, this is not reflected in the rank scores presented which are rounded to the first decimal place. One of the major general complaints against ARS is the wide variance seen by different ARS for lower ranked schools. Going back to the USNWR study, this means for the schools with lower rankings that received say 50 ranked scores, for the ratings to be accurate within 0.034, the total sum of all the 50 rankings would have to be within 2 of the true sum. Since it is not reasonable to assume that this type of accuracy is possible, either more reliable surveys are necessary or perhaps the rankings should be expressed at a lower level of resolution, like quartiles. The American Mathematical Society, for example, groups math doctoral programs in this manner into Group I, II, III, and IV schools.

The introduction of a new way to test the validity of a GPARS also raises some theoretical questions. Hypothesis tests of the form proposed in this thesis usually look at the difference in the value of a regression parameter between two populations. But the hypothesis test used in this thesis

is of a slightly different nature. The populations considered are the same but the set of regressors considered are different. The theoretical basis for a formal hypothesis test of this nature, not to mention the model assumptions that have to be met, could not be found. The use of confidence intervals also assumes a random sample, which coming from a survey it is most certainly not. It is also important to note that a more realistic hypothesis for the halo effect would take into account it is not an "all or none" effect. In other words, the reputation of an institution may only partially influence the rating. This fact is important because it affects our interpretation of the results. Note a null result indicates under the most extreme scenario (i.e. every respondent either makes their decision either solely based on only the halo effect or just the institutional characteristics) we see no evidence for the halo effect. Hence, a null result is a strong result (since under the most extreme case no difference is observed). However, a result that favors the alternative hypothesis is not as strong since in reality we know the halo effect is not as extreme, so it is uncertain if this result would hold in more realistic scenario. Hence, even if the null hypothesis were rejected, more research would have to be done to devise better models of the halo effect. Such research would have to determine fundamental issues such as whether the halo effect is a bias that affects a portion of the respondents or is a systemic bias the affects all respondents (to varying degrees).

4.3 Suggestions for Future Research

Suggestions for future research would include the use of more covariates, covariates that are often underappreciated (such as contact hours professors spend with students), different metrics to measure a given quality, a closer examination into the theoretical basis for the proposed hypothesis test and ways to model the halo effect, and perhaps paying more attention to the quality of data collected. A more thorough study of the nature and reliability of reputational surveys

themselves (something along the lines of the "R-rank" and "S-rank" surveys used by the NRC, but perhaps using more complete and in-depth in person interviews) might help also devise surveys that provided better measures of academic quality. Finally, ranking systems that did not emphasize computing quantitative scores but rather grouped schools into classes might be more realistic, accurate, and useful to potential students.

In principle, this methodology could be extended to other doctoral graduate programs (with the potential caveat that the nature of the surveys and covariates may have to be adapted and modified for different disciplines). It would also be nice to use the results from the list of ASA schools to extend the scope of this ranking system to more statistics doctoral programs. Finally, considering the wide range of Master's programs and the fact that the ASA published a white paper in 2012 on recommended requirements a master's statistics program should meet (allowing a defined list to be constructed), perhaps this ranking methodology could be extended to Master's programs in statistics.

SECTION 5 CONCLUSION

A number of serious issues plagues academic ranking systems such as those published by U.S. News and World Report. The use of reputational surveys appears to address most of these issues except validity. This thesis considered a major potential issue affecting the validity of reputational survey based academic ranking systems for graduate programs called the "halo effect". In the context of this study, the halo effect would cause some respondents to base their ratings partially on the prestige or reputation of the home institution and not solely on the merits or quality of the graduate program considered. A statistical hypothesis test for the halo effect was proposed based on the premise that evidence against the halo effect would be a model based purely on department specific measures of academic quality or performance that was better than any model based only on non-program specific metrics of academic performance. This hypothesis test was applied to the 2012 USNWR rankings of U.S. statistics graduate programs using metrics proposed in previous studies of graduate program academic ranking systems ("Preparing Master's Statistics", 2016). Although the best model based only on program specific metrics was better than the best model based on non-program specific metrics, the improvement was not statistically significant. Owing to the limited number of covariates considered, more comprehensive tests might be needed to obtain more reliable results. In addition, the statistical hypothesis test might require some modification to take into account that non-program specific metrics only have a partial effect on the ratings.

REFERENCES

Academic Ranking of World Universities. (n.d.). Retrieved from https://en.wikipedia.org/wiki/Academic_Ranking_of_World_Universities#Methodology

Adler, N.J. and Harzing, A. (2009). When Knowledge Wins: Transcending the Sense and Nonsense of Academic Ranking. *Academy of Management Learning & Education* 8(1), 72–95.

Aguillo, I.F., Granadino, B., Ortega, J.L. & Prieto, J.A. (2006). Scientific research activity and communication measured with cybermetrics indicators. *Journal of the American Society for Information Science and Technology*, **57**(10), 1296-1302.

Astin, A. W. (1985). *Achieving educational excellence*. San Francisco: Jossey-Bass.

Best Statistics Programs | Top Statistics Schools | US News Best Graduate Schools. (n.d.). Retrieved June 06, 2016, from http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-science-schools/statistics-rankings

Brennan, J., Brodnick, R., and Pinckley , D. (2008). De-Mystifying the U.S. News Rankings: How to Understand What Matters, What Doesn't and What You can Actually Do About It, *Journal of Marketing for Higher Education*, 17(2): 338-342.

Cartter, A. A. (1966). *An assessment of quality in graduate education*. Washington, D. C.: American Council on Education.

Cattell, J. M., & Cattell, J. (1906). *American men and women of science: 1st-13th ed.* Lancaster, Pa.: Science Press.

Center for World University Rankings. (n.d.). Retrieved from http://cwur.org/methodology/

CWTS Leiden Rankings: Indicators. (n.d.). Retrieved from http://www.leidenranking.com/information/indicators

Clark, K. E. (1957). *America's psychologists: a survey of a growing profession*. Washington, D.C: American Psychological Association.

Clark, M. J. (1974). Dimensions of quality in doctoral education. *Findings*, 1(4), 1-4.

College and university rankings. (n.d.). Retrieved June 04, 2016, from https://en.wikipedia.org/wiki/College_and_university_rankings

Davis, H.D. and Diamond, N., (1997). *The Rise of American Research Universities: Elites and Challengers in the Postwar Era*. Baltimore: Johns Hopkins University Press, 1997

Diver, C. (2005, November). Is there Life After Rankings? *The Atlantic* . A1

Grunig, S. D. (1997). Research, reputation, and resources: The effect of research activity on perceptions of undergraduate education and institutional resource acquisition. *The Journal of Higher Education, 68*(1), 17-52.

How U.S. News Calculated the Best Global Universities Rankings. (n.d.). Retrieved June 04, 2016, from http://www.usnews.com/education/best-global-universities/articles/methodology

Hughes, R. (1925). *A Study of Graduate Schools of America.* Oxford, OH: Miami University Press.

Keniston, H. (1959). *Graduate study and research in the arts*. Philadelphia: University of Pennsylvania Press.

King, S., and Wolfle, L. (1987). A latent- variable causal model of faculty reputational ratings. *Research in Higher Education* 27(2): 99-106.

Kivinen, O., & Hedman, J. (2008). World-wide university rankings: A Scandinavian approach. *Scientometrics, 74*(3), 391-408.

Lawrence, J. K., & Green, K. C. (1980). *A question of quality. The higher education r ratings game*. AAHE-ERIC/Higher Education Research Report No. 5. Washington, DC: American Association for Higher Education.

Liu, N. C., Cheng, Y. (2005), The academic ranking of world universities, *Higher Education in Europe*, 30 (2) : 127–136.

Methodology: Best Science Schools Rankings. (n.d.). Retrieved June 06, 2016, from http://www.usnews.com/education/best-graduate-schools/articles/science-schools-methodology

National Science Foundation, National Science Board. (1969). *Graduate education: parameters for public policy*. Washington, D. C.

Preparing Master's Statistics Students for Success: A Perspective from Recent Graduates and Employers. (n.d.). Retrieved June 04, 2016, from http://magazine.amstat.org/blog/2013/02/01/mastersworkgrou/

Saunier, M. (1985). Objective measures as predictors of reputational ratings. *Research in Higher Education* 23(3): 227-244.

Shin, J. C., Toutkoushian, R. K., & Teichler, U. (2011). *University rankings : Theoretical basis, methodology and impacts on global higher education*. Dordrecht: Springer

Skapinker, M. (2008, January 7). Why business ignores business schools. *Financial Times*, (18) 18-27.

Stecklow, S. (1995, April 5). Cheat Sheets: Colleges Inflate SATs And Graduation Rates In Popular Guidebooks Schools Say They Must Fib To U.S. News and Others To Compete Effectively. Moody's Requires the Truth. *Wall Street Journal*, A1.

Webster, David S., Clifton F. Conrad, and Eric L. Jensen. (1991). "Objective and reputational rankings of Ph. D.-granting departments of sociology, 1965–1982." *Sociological Focus* 21(2). 177-198.

Zhang, G. and Chen, J. J., (2015). Biostatistics Faculty and NIH Awards at U.S. Medical Schools. *The American Statistician*, 69(1), 34-40.

Appendix 1 List of Schools Eligible to Be Ranked (Compiled by ASA)

University	Department
Arizona State University	School of Mathematical and Statistical Sciences
Baylor University	Department of Statistical Science
Boston University School of Public Health (joint w/ Grad School Dept Math & Stats)	Department of Biostatistics
Brown University School of Public Health	Department of Biostatistics
Carnegie Mellon University	Department of Statistics
Case Western Reserve University	Department of Epidemiology and Biostatistics
Case Western Reserve University	Department of Statistics
Colorado State University	Department of Statistics
Columbia University	Department of Statistics
Columbia University, Mailman School of Public Health	Department of Biostatistics
Cornell University	Department of Statistical Science
Duke University	Department of Statistical Science
Emory University Rollins School of Public Health	Department of Biostatistics and Bioinformatics
Florida State University	Department of Statistics and Statistical Consulting Center
George Washington University	Department of Statistics
Harvard University	Department of Biostatistics
Harvard University	Department of Statistics
Iowa State University	Department of Statistics & Statistical Laboratory
Johns Hopkins Bloomberg School of Public Health	Department of Biostatistics
Kansas State University	Department of Statistics
Medical College of Wisconsin	Division of Biostatistics
Medical University of South Carolina, Department of Public Health Sciences	Division of Biostatistics and Epidemiology
Michigan State University	Department of Statistics and Probability
New York University Stern School of Business	Department of Information, Operations and Management Science
North Carolina State University	Department of Statistics
North Dakota State University	Department of Statistics
Northwestern University	Department of Statistics
Oklahoma State University	Department of Statistics
Oregon State University	Department of Statistics
Penn State University	Department of Statistics
Purdue University	Department of Statistics
Rice University	Department of Statistics
Rutgers University	Department of Statistics and Biostatistics
Southern Methodist University	Department of Statistical Science
Stanford University	Department of Statistics
Temple University Fox School of Business	Department of Statistics
Texas A&M University	Department of Statistics

The Ohio State University	Department of Statistics
The University of Alabama	Information Systems, Statistics and Management Science Department
Tulane School of Public Health and Tropical Medicine	Department of Biostatistics and Bioinformatics
University at Albany School of Public Health	Department of Epidemiology and Biostatistics
University at Buffalo	Department of Biostatistics
University of Louisville School of Public Health and Information Sciences	Department of Bioinformatics and Biostatistics
University of Alabama Birmingham	Department of Biostatistics
University of California Berkeley	Graduate Group in Biostatistics
University of California Berkeley	Department of Statistics
University of California Davis	Department of Statistics
University of California Los Angeles	Department of Statistics
University of California Los Angeles Fielding School of Public Health	Department of Biostatistics
University of California Riverside	Department of Statistics
University of California Santa Barbara	Department of Statistics and Applied Probability
University of Chicago	Department of Statistics
University of Cincinnati College of Medicine, Department of Environmental Health	Division of Epidemiology and Biostatistics
University of Colorado Denver	Department of Mathematical and Statistical Sciences
University of Connecticut	Department of Statistics
University of Florida	Department of Statistics
University of Georgia	Department of Statistics
University of Illinois at Chicago School of Public Health	Epidemiology and Biostatistics Division
University of Illinois at Urbana-Champaign	Department of Statistics
University of Iowa	Department of Statistics and Actuarial Science
University of Iowa College of Public Health	Department of Biostatistics
University of Kentucky	Department of Statistics
University of Michigan	Department of Biostatistics
University of Michigan	Department of Statistics
University of Minnesota	School of Statistics
University of Minnesota School of Public Health	Division of Biostatistics
University of Missouri	Department of Statistics
University of Nebraska	Department of Statistics
University of North Carolina at Chapel Hill	Department of Biostatistics
University of North Carolina at Chapel Hill	Department of Statistics and Operations Research
University of Pennsylvania	Department of Statistics
University of Pennsylvania Perelman School of Medicine	Department of Biostatistics and Epidemiology
University of Pittsburgh	Department of Statistics
University of Pittsburgh Graduate School of Public Health	Department of Biostatistics
University of Rochester Medical Center	Department of Biostatistics and Computational Biology
University of South Carolina	Department of Statistics

University of South Carolina Arnold School of Public Health	Department of Epidemiology and Biostatistics
University of Texas School of Public Health	Division of Biostatistics
University of Virginia	Department of Statistics
University of Washington	Department of Biostatistics
University of Washington	Department of Statistics
University of Wisconsin	Department of Statistics
Virginia Commonwealth University School of Medicine	Department of Biostatistics
Virginia Tech	Department of Statistics
Western Michigan University	Department of Statistics
Yale School of Public Health	Department of Biostatistics
Yale University	Department of Statistics

Appendix 2 Copy of Survey Used by USNWR (first page)

Job No. 13-075936-01
Card 01
ID No. (1-6)
Version 59 (7)
IID No. _____ (8-17)

Best Graduate Schools
2014 Peer Assessment of Doctoral Programs in Statistics

As part of its spring 2014 report on graduate and professional schools, U.S. News & World Report is conducting a survey of doctoral programs in statistics. The universe of programs was developed in conjunction with the American Statistical Association. They are listed below by state. This survey is being sent to the chair of the statistics or biostatistics department and the director of graduate studies (or alternatively, a senior faculty member who teaches graduate students) in each of these departments. **Your participation in this survey is greatly appreciated.**

Directions for Rating Doctoral Statistics Programs:

1. Please rate the academic quality of the doctoral statistics program at each school with which you are familiar. Consider all factors that bear on or give evidence of the excellence of the school's doctoral statistics program, for example, curriculum, record of scholarship, quality of faculty and graduates.
2. You are encouraged to review the entire list first, before beginning to rate individual programs.
3. Using a **black** pen, rate each school on a scale of marginal (1) to outstanding (5) by marking an "X" in the corresponding box. If you are not familiar with a school's program, please put an "X" in the box labeled "Don't Know".
4. Return the completed survey in the postage-paid return envelope provided by **December 13, 2013**. If you have misplaced the envelope, return the survey to: Angela Foster-Woods, Ipsos, 222 South Riverside Plaza Chicago, IL 60606-5809.
5. Keep a copy of your completed survey for your own records.

Any questions? Contact the *Best Graduate Schools*, help desk at (800) 408-5365 or usnews@ipsos.com. Thank you!

		RATINGS						
		Outstanding 5	Strong 4	Good 3	Adequate 2	Marginal 1	Don't Know DK	
	ALABAMA							
59001	University of Alabama	☐	☐	☐	☐	☐	☐	(21)
59076	*University of Alabama – Birmingham	☐	☐	☐	☐	☐	☐	
	ARIZONA							
59077	Arizona State University	☐	☐	☐	☐	☐	☐	
	CALIFORNIA							
59002	Stanford University	☐	☐	☐	☐	☐	☐	
59003	University of California – Berkeley	☐	☐	☐	☐	☐	☐	
59004	*University of California – Berkeley	☐	☐	☐	☐	☐	☐	
59005	University of California – Davis	☐	☐	☐	☐	☐	☐	
59006	University of California – Los Angeles	☐	☐	☐	☐	☐	☐	
59007	*University of California – Los Angeles	☐	☐	☐	☐	☐	☐	
59008	University of California – Riverside	☐	☐	☐	☐	☐	☐	
59009	University of California – Santa Barbara	☐	☐	☐	☐	☐	☐	
	COLORADO							
59010	Colorado State University	☐	☐	☐	☐	☐	☐	
59078	University of Colorado – Denver	☐	☐	☐	☐	☐	☐	(33)

Schools marked with an asterisk (*) denote a program in the department of biostatistics.

V59-1

Appendix 3 Published 2013 U.S. Statistics Graduate Program Rankings by USNWR

USNWR Rank Score	University	Department
4.9	Stanford University	Statistics
4.7	University of California Berkeley	Statistics
4.6	Harvard University	Biostatistics
4.6	University of Washington	Biostatistics
4.4	Johns Hopkins	Biostatistics
4.4	University of Chicago	Statistics
4.3	Harvard University	Statistics
4.3	University of Washington	Statistics
4.2	Carnegie Mellon University	Statistics
4.1	Duke University	Statistics
4.1	University of Pennsylvania	Statistics
4	University of Michigan	Biostatistics
4	University of North Carolina at Chapel Hill	Biostatistics
4	University of Wisconsin	Statistics
3.9	North Carolina State University	Statistics
3.9	Texas A&M University	Statistics
3.9	University of California Berkeley	Biostatistics
3.9	University of Michigan	Statistics
3.8	Iowa State University	Statistics
3.7	Columbia University	Statistics
3.7	Penn State University	Statistics
3.7	University of Minnesota	Statistics
3.7	University of North Carolina at Chapel Hill	Statistics & OR
3.6	Cornell University	Statistics
3.6	Purdue University	Statistics
3.6	University of Minnesota	Biostatistics
3.5	The Ohio State University	Statistics
3.5	University of California Davis	Statistics
3.4	Columbia University	Biostatistics
3.4	University of California Los Angeles	Statistics
3.4	University of California Los Angeles	Biostatistics
3.4	University of Florida	Statistics
3.3	University of Illinois at Urbana-Champaign	Statistics
3.3	University of Iowa	Statistics
3.3	Yale School of Public Health	Biostatistics
3.3	Yale University	Statistics
3.2	Emory University	Biostatistics
3.1	Florida State University	Statistics
3.1	Rice University	Statistics

3.1	Rutgers University	Biostatistics
3	Brown University	Biostatistics
3	Colorado State University	Statistics
3	University of Connecticut	Statistics
3	University of Iowa	Biostatistics
2.9	Michigan State University	Statistics
2.9	University of Pittsburgh	Biostatistics
2.8	Boston University	Biostatistics
2.8	Northwestern University	Statistics
2.8	University of Pittsburgh	Statistics
2.7	George Washington University	Statistics
2.7	University of Georgia	Statistics
2.7	University of Illinois at Chicago	Biostatistics
2.7	University of Missouri	Statistics
2.7	Virginia Tech	Statistics
2.6	Southern Methodist University	Statistics
2.5	University of California Santa Barbara	Statistics
2.4	Arizona State University	Statistics
2.4	Oregon State University	Statistics
2.4	University of South Carolina	Statistics
2.4	University of Virginia	Statistics
2.3	Temple University Fox School of Business	Statistics
2.3	University of California Riverside	Statistics
2.2	The University of Alabama	Biostatistics
2.2	University at Albany	Biostatistics
2.2	University at Buffalo	Biostatistics
2.2	University of Colorado Denver	Statistics
2.1	Baylor University	Statistics
2.1	University of Kentucky	Statistics
2	University of South Carolina	Biostatistics
3.5	Univ of Pennsylvania	Biostatistics
3	University of Rochester	Biostatistics
2.8	NYU (Stern)	Business
2.7	Medical College of Wisconsin	Biostatistics
2.7	Univ Texas Health Sci Center - Houston	Biostatistics
2.6	CWRU	Biostatistics
2.3	Medical University of South Carolina	Biostatistics
2.2	Kansas State University	Statistics
2.1	Case Western Reserve University	Statistics
2	Virginia Commonwealth Univ	Biostatistics